Nordic Yearbook of Law and Informatics 2020–2021

Law in the Era of Artificial Intelligence

The Swedish Law and Informatics Research Institute

Liane Colonna & Stanley Greenstein (editors)

NORDIC YEARBOOK OF LAW AND INFORMATICS 2020–2021

Nordic Yearbook of Law and Informatics 2020–2021
Law in the Era of Artificial Intelligence
Liane Colonna & Stanley Greenstein (eds)

Distributor
eddy.se ab
Box 1310
SE-621 24 Visby, Sweden
www.bokorder.se
order@bokorder.se

Produced by Stiftelsen Juridisk Fakultetslitteratur (SJF)
and The Swedish Law and Informatics Research Institute (IRI)
Law Faculty, Stockholm University
SE-106 91 Stockholm, Sweden
irilaw.org
iri@juridicum.su.se

ISBN 978-91-8892-964-8

Nordic Yearbook of Law and Informatics 2020–2021

Law in the Era of Artificial Intelligence

Edited by

Liane Colonna
Stanley Greenstein

Foreword

The 35th Nordic Conference on Law and Information Technology was held in Stockholm, 11–12 November 2020. As on previous occasions, *The Swedish Law and Informatics Research Institute* (IRI) had the privilege of arranging the conference in conjunction with *The Foundation for Legal Information* (Stiftelsen för rättsinformation) and *The Swedish Society for IT and Law* (Svenska föreningen för IT & Juridik) SIJU.

Preparations for the conference were hampered by the continually changing restrictions and guidelines brought on by the Covid-19 pandemic that was rampant at the time. To comply with these restrictions yet still host the conference, the decision was made to hold the conference with a hybrid format, resulting in most of the participants attending the conference online with just a handful of persons representing the organizers and a few participants on site. While this format reduced the 'in real life' social interaction that is so much a characteristic of the Nordic Conference on Law and IT, it increased its accessibility for many, and the conference can only be described as a great success despite challenging circumstances.

The overall title of the conference was *Law in the Era of Artificial Intelligence*. The main theme was how an increased use of Artificial Intelligence (AI) is influencing the meaning of previously established legal concepts not yet adapted to a society increasingly reliant on AI. In this respect, the conference was divided up into four sessions each focusing on a selected aspect, namely, session 1 on *Data Protection*, session 2 on *Transparency*, session 3 on *Liability* and finally session 4 on *Regulation*. The sessions began with the presentations of speakers with different backgrounds and extensive experience concerning the addressed issues. Several international experts also participated. In

addition to drawing a picture of where we stand today, each session included a moderated discussion.

Like the conference, the Nordic Yearbook of Law and Informatics 2020–2021 is entitled *Law in the Era of Artificial Intelligence* and as per tradition it is made up of contributions from speakers that participated in the conference.[1] Consequently, this Nordic Yearbook of Law and Informatics is divided in to four parts, which reflects the respective sessions that made up the 35th Nordic Conference:

Data protection: How is data protection and privacy affected? AI is in crucial parts a data-driven phenomenon. The amount of data increases as systems develop: possible gains in efficiency and quality are substantial. Simultaneously, it is obvious that security risks, sources of errors and opportunities for misuse are considerable. Which are the real risks and what is the role of law?

Transparency: How is transparency ensured? AI applications can function autonomously and change behaviors over time without human interaction. How can independent and dynamic processes be understood and controlled?

Liability: How should responsibility be allocated? AI is often integrated as a component in complex systems of systems. To what extent is it possible to investigate causality and upon whom can responsibility be assigned for errors and accidents? Also, to what extent do traditional legal concept remain relevant?

Regulation: How can AI be regulated? To function in a digital society laws and other regulations must frequently be built-in into the systems. This movement is sometimes referred to as 'code as law' and is closely connected to 'techno regulation' and 'value sensitive design'. For example, law must be embedded into the system design of self-driving vehicles so that they can adhere to the traffic rules. How can the law be developed to meet the need of proactive and operative forms of rules?

Finally, mention should be made of the fact that many articles in the Nordic Yearbook are based on presentations made at the 35th Nordic Conference, and thus were completed prior to the publication of the draft of the AI Act.[2]

1 The full program is available at https://irilaw.org/e20/program/ and a presentation of the authors can be found after the articles.

2 European Commission, *Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonized Rules on Artificial Intelligence (Artificial Intelli-*

***

The 35th Nordic Conference on Law and Information Technology and this publication have been made possible by financial support generously provided by *Justitierådet Edvard Cassels stiftelse,* SIJU and the *Law Faculty's Trust Fund for Publications, Stockholm University* (Stiftelsen Juridisk Fakultetslitteratur). Meticulous and thorough work with the editing of the yearbook was provided by *Linnéa Holmén* and *Ebony A. Wade*. Special thanks to *Stiftelsen Juridisk Fakultetslitteratur* (SJF) who financed the electronic version of the Nordic Yearbook of Law and Informatics 2020–2021.

We also give thanks to all the contributors to the Nordic Yearbook 2020–2021 who have produced excellent articles under tight deadlines. We are convinced that the readers of this Nordic Yearbook will benefit from the wealth of knowledge embedded in these articles.

Stockholm in December 2021

Liane Colonna *Stanley Greenstein*

gence Act) and Amending Certain European Union Acts, Brussels, 21.4.2021 COM(2021) 206 final, available at https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX-%3A52021PC0206.

Table of Contents

Part 4: How to Regulate?

Part 1

Data Protection

Legal Implications of Using AI as an Exam Invigilator*

LIANE COLONNA

The Covid-19 pandemic has taken the world by surprise, forcing many countries to adopt shelter-in-place directives, as well as partial or total lockdown and social distancing orders in order to contain the spread of the virus. Universities around the globe have been profoundly affected by stay-at-home orders, which have required them to close their doors and shift to online education. Despite long-standing skepticism to online teaching and learning, at least compared to active and in-person education, it has become the main platform for education during the pandemic, creating colossal pedagogical changes. One of the biggest challenges that universities have had to confront due to the unexpected and sudden shift to online education concerns what kind of assessment techniques are appropriate in an online environment.

In an effort to avoid delaying or postponing examinations amid the Covid-19 outbreak, many universities have turned to online proctoring tools, raising complex questions about how they can ensure the integrity of online assessments while at the same time respect ethical and legal constraints, especially regarding students' fundamental rights to privacy, data protection and non-discrimination.[1] While institutions insist that these tools are necessary in

* The support of The Wallenberg AI, Autonomous Systems and Software Program – Humanities and Society (WASP-HS), Ethical and Legal Challenges in Relationship to AI-driven Practices in Higher Education (MMW2020.0138), is gratefully acknowledged.

1 *See* Neil Selwyn et al., *A Necessary Evil? The Rise of Online Exam Proctoring in Australian Universities*, Media Int'l Australia (2021).

order to fulfil the requirements of distance education and to ensure the integrity of the exams, students raise legitimate concerns about whether universities have lawful grounds to process their personal data, particularly when their consent is not provided. They also raise questions about the surveillance effect of online proctoring, which can increase testing anxiety as well as diminish trust and cooperation between students and institutions. Students are further concerned about the technical and social biases that can be embedded into the algorithms that fuel the technology, leading to marginalized students disproportionately and unfairly having to pay the price of these technologies, based on potentially racist, sexist, ableist, and hetero-centrist norms being reflected in the systems.

This article considers the legal implications of the use of remote proctoring using artificial intelligence (AI) to monitor online exams and, in particular, to validate students' identities and to flag suspicious activities during the exam to discourage academic misconduct like plagiarism, unauthorized collaboration and sharing of test questions or answers. The emphasis is on AI-based facial recognition technologies (FRT) that can be used during the authentication process for remote users during the online exam process as well as to identify dubious behavior throughout the examination. The central question explored is whether remote proctoring systems are necessary and lawful based on European human rights law.

The first part of the paper explores the use of AI-based remote proctoring technologies in higher education (HE), both from the institutional perspective as well as from the student perspective. It emphasizes how universities are shifting from a reliance on systems that include human oversight, like proctors overseeing the examinations from remote locations, towards more algorithmically driven practices that rely on processing biometric data. The second part of the paper examines how the use of AI-based remote proctoring technologies in HE impacts the fundamental rights of students, focusing on the fundamental rights to privacy, data protection, and non-discrimination. Next, it provides a brief overview of the legal frameworks that exists to limit the use of this technology. Finally, the paper closely examines the issue of legality of processing in an effort to unpack and understand the complex legal and ethical issues that arise in this context.

Online proctoring tools

In a conventional classroom, exams are proctored by a human being who monitors the students and the physical environment during the exam. In the context of online exams, there is likewise a need for reliable and inexpensive monitoring abilities in order to authenticate test takers' identities, observe the test taker's behavior to preserve academic integrity, and secure test content.[2] Currently, there exist different methods for online proctoring. The focus herein is on live proctored testing and AI proctored testing.

The first group of methods rely on human proctors watching the test takers through a webcam from a remote location.[3] Sometimes this approach is referred to as "online human monitoring."[4] Typically, the online proctoring process starts with a verification of the exam-taker's identity.[5] This may happen by, for example, presenting an identity card like a student card, a driving license or passport to the proctor via the student's webcam.[6] Next, the proctor may ask each test-taker to move his or her webcam around to scan their physical testing environment in order to make sure the student does not have access to unpermitted items like phones or books.[7] It is also important to note that key functionalities from the test-taker's computer may be disabled, like copying, pasting, printing, taking a screen shot or accessing other applications.[8] During the exam, the proctors watch and listen for any unusual behaviors of the test taker, such as unusual eye movements or removing oneself from the field of view, and can alert the test taker or even stop the test in the event

2 Thomas Langenfeld, *Internet-Based Proctored Assessment: Security and Fairness Issues*, 39 Educational Measurement: Issues & Practice 24 (2020).

3 Yousef Atoum et al., *Automated Online Exam Proctoring*, 19 IEEE Transactions on Multimedia 1609 (2017).

4 *Id.*

5 Aiman A. Turani, Jawad H. Alkhateeb & AbdulRahman A. Alsewari, *Students Online Exam Proctoring: A Case Study Using 360 Degree Security Cameras*, 2020 Emerging Technology in Computing, Communication and Electronics (ETCCE), 1–5 (2020).

6 *Id.*

7 *Id.*

8 *Id.*

of suspicious behavior.[9] These types of proctoring exams can take place in real time or, for a more flexible model, can be pre-recorded and played back to a proctor at a later point in time.[10] They can be given at local testing centers but due to the increased ownership of laptops and tablet computers, and, of course, the pandemic situation, they have been increasingly administered in students' homes.[11]

Drawbacks to this approach include labor intensiveness and cost as it often takes many human proctors to monitor the test-takers.[12] Additionally, the proctor might have limited vision and may not be able to observe all cheating strategies, such as notes laying on a test taker's desk.[13] As noted, it may be possible for the remote proctor to ask the test-taker to sweep the room using his or her webcam, but this may create undue pressure and stress for the test taker, as well as reveal intimate information about the student's private life to the remote, human proctor.[14] Furthermore, remote proctored exams require well-established infrastructure, on both the student and institutional sides, including software, hardware and a stable internet connection.[15]

Because of these drawbacks, vendors have begun to supplement live proctoring with AI proctoring, which can automatically detect indications of possible fraud. Since AI is an umbrella term, denoting the use of many different types of technologies, it is important to clarify at the outset that this article is focused on the use of FRT, an application of computer vision. FRT is a "touchless" form of biometric that makes it possible to track an individual based on, for example, iris recognition and facial recognition.[16] More specifically, an algorithm is used to recognize a human face through the use

9 Atoum et al., *supra* note 3.

10 Turani, Alkhateeb & Alsewari, *supra* note 5.

11 Selwyn, *supra* note 1.

12 Atoum et al., *supra* note 3.

13 *Id.*

14 *Id.*

15 Fiseha M. Guangul et al., *Challenges of Remote Assessment in Higher Education in the Context of COVID-19: A Case Study of Middle East College*, 32 Educational Assessment, Evaluation and Accountability 519 (2020).

16 Jeff D. Neuburger, *Will the Role of Facial Recognition Grow In a Post-COVID-19 World?* 25 Cyberspace Lawyer 4 (2020).

of biometrics, which track facial features from a photo or video.[17] These facial features often include the distance between a person's eyes, the distance from their forehead to their chin, and other "facial landmarks."[18]

FRT can be used to support (or replace) human proctors in a number of different ways. First, it can be used to recognize students' in-test (mis)behavior by checking room conditions and analyzing behavior that might indicate cheating.[19] Furthermore, FRT can identify additional faces in a given testing environment that may be assisting the student in an inappropriate manner.[20] It can also recognize unauthorized objects in a testing environment.[21] Furthermore, it can track eye movements which may indicate misconduct, like looking away from the screen.[22]

Besides helping to ensure academic integrity, FRT has a key role to play regarding authentication in advancing proctoring methods.[23] Biometric authentication methodologies rely on intrinsic physical and behavioral traits rather than things like username/passwords or access card/PINs.[24] For example, in biometric verification, discussed more below, FRT might be used to match a student's photographed ID with the student's facial features.[25] Thereafter, biometric data can be captured continuously from the user during an exam session in

17 Steve Symanovich, *How Does Facial Recognition Work?*, Norton (February 8, 2019) http://us.norton.com/internetsecurity-iot-how-facial-recognition-software-works.html (last accessed Apr. 27, 2021).

18 *Id.*

19 Evan Selinger, *Abolish A.I. Proctoring*, https://onezero.medium.com/abolish-a-i-proctoring-c9e017dd764f (last accessed April 27, 2021); Selwyn, *supra* note 1.

20 Aileen Scott, *Artificial Intelligence Is Making Online Proctoring Safe and Secure*, Medium (March 14, 2019), http://medium.com/@aileenscott604/artificial-intelligence-is-making-online-proctoring-safe-and-secure-9b03845602da (last accessed April 27, 2021).

21 *Id.*

22 Selinger, *supra* note 19; Scott, *supra* note 20.

23 Atoum et al., *supra* note 3.

24 Corey Ashby, Amit Bhatia, Francesco Tenore, Jacob Vogelstein, *Low-cost Electroencephalogram (EEG) Based Authentication*, 5th International IEEE/EMBS Conference on Neural Engineering 442–445 (2011).

25 Selwyn, *supra* note 1.

order to verify the exam taker throughout the session.[26] This might be accomplished through, for example, the use of eye tracking and facial detection.[27]

Fundamental rights implications of proctoring exams

Privacy and data protection

Privacy and data protection have long been leading concerns with eLearning, and when it comes to online proctoring systems the issues only multiply.[28] A first concern is that of the legality of processing, especially concerning the processing of biometric data like student' faces as well as student's living spaces.[29] This issue will be explored in detail below.

A related, and equally complex, concern involves the role of automated individual decision making, including profiling, where there is no human involvement. Here, concerns arise where AI flags a student for cheating when the behavior is not actionable; for example, where a student's child enters the test-taker's environment to ask for a snack, causing the student to look away from the screen towards a second person. Other scenarios can easily be imagined such as where a student urgently needs to get up to urinate or change a sanitary pad.[30] While many software providers insist that their proctoring systems are trustworthy to the extent that humans can review the computer-generated results before sanctions are imposed, it is not clear that ex post human review is capable of mitigating the risks of

26 Hadian S. G. Asep & Y. Bandung, *A Design of Continuous User Verification for Online Exam Proctoring on M-Learning*, 2019 International Conference on Electrical Engineering and Informatics 284–289 (2019).

27 Selwyn, *supra* note 1.

28 Faten F. Kharbat & Ajayeb S. Abu Daabes, *E-proctored Exams During the COVID-19 Pandemic: A Close Understanding*, Education and Information Technologies (2021).

29 Anandi Barker, *Big Brother is Proctoring You*, The Daily Texan (September 23, 2020), http://thedailytexan.com/2020/09/23/big-brother-is-proctoring-you/ (last accessed April 27, 2021).

30 Heather Murphy, *She Was Going Into Labor. But She Had a Bar Exam to Finish*. The New York Times (September 13, 2020)(last accessed May 3, 2021)(explaining how students have urinated in their seats in order to avoid being flagged for cheating and how one woman even gave birth during a remote exam).

harms created by these systems, particularly as the exam review process can be expensive, complex and outsourced to actors far removed from the local context.[31]

It is also important to mention the surveillance effect that may arise when students feel as though they are being watched under a constant microscope as they take online exams in their homes, traditionally an area designated with a very high level of privacy protection.[32] Not only can intensely personal information about a student, such as their lifestyle choices and socio-economic status, be revealed in the home setting, but online assessment tools may also make students feel like cheaters, even before submitting any work.[33] The surveillance capabilities of AI-based proctoring tools may create a lack of trust and cooperation between the students and the institution by causing them to feel that they are in a "less nurturing, comfortable learning environment."[34] It may also exacerbate existing test anxiety[35] causing some students to refrain from taking certain exams out of fear that they would be accused of cheating for accidentally moving

31 Michael Dodge, *Online Exam Monitoring is Now Common in Australian Universities — But Is It Here to Stay?* The Conversation (April 18, 2021), http://theconversation.com/online-exam-monitoring-is-now-common-in-australian-universities-but-is-it-here-to-stay-159074 (last accessed April 26, 2021)(explaining that suspicious behaviour can be reviewed by a "live" remote proctor. This work is often outsourced to developing nations such as India and the Philippines, where remote proctors are reportedly paid around $3.50 per hour.)

32 Beverly Balos, *A Man's Home Is His Castle: How the Law Shelters Domestic Violence and Sexual Harassment*, 23 St. Louis U. Pub. L. Rev. 77, 90 (2004)("The home and non-interference with the sanctity of home is well established. It is not just a physical place but is imbued with idealized characteristics. It is a place of respite from the commercial marketplace. It fosters intimate relationships and allows family life to flourish. It is also a place of safety and physical comfort. Beyond relational intimacy, the home also functions as a symbol for a feeling of belonging and a place where one can realize one's potential.")

33 Jessica Wong, *Post-secondary Students Call for Changes to Online Exam Rules as Cheating Concerns Rise*, CBC News (October 25, 2020), http://www.cbc.ca/news/canada/post-secondary-assessment-integrity-proctoring-1.5767953 (last accessed April 27, 2021).

34 Alisia LoSardo, *Faceoff: The Fight for Privacy in American Public Schools in the Wake of Facial Recognition Technology*, 44 Seton Hall Legis. J. 373, 383–87 (2020); J. William Tucker & Amelia Vance, *School Surveillance: The Consequences for Equity and Privacy*, 2 Education Leaders Report 4, 8 (2016).

35 Barker, *supra* note 29.

too much or going off screen, particularly those with disabilities who do not want to (or have the means to) go through the "grueling, exposing and expensive process" of requesting accommodations.[36] Essentially, students are required to show their private homes, be in an interruption-free space with sufficient lighting, have a computer and stable internet connection, maintain consistent eye contact with a webcam — in addition to knowing the course content, which understandably can seriously increase their anxiety and break down trust.[37]

Some experts argue that FRT will breed a "generation that will be comfortable with and fully accepting of total government surveillance."[38] In other words, FRT might "normalize invasive means of surveillance in the eyes of students."[39] One commentator notes: "When professors rely on proctoring services, they devalue their students' privacy and mental ease while forcing them to demonstrate their comprehension of class material in almost dystopian conditions."[40] In short, there is a genuine concern that "surveillance pedagogy" is becoming entrenched in contemporary education.[41]

Relatedly, the use of FRT may create a threat to intellectual privacy, understood as "a much-needed protection for learning, reading and communicating" that helps students develop their free thoughts, creativity, risk-taking and overall inquisitiveness.[42] Tucker and Vance explain: "If students feel as though they cannot step outside of the mainstream for fear of ridicule or are afraid to ask a question because their ignorance might be captured forever in the virtual cloud, then surveillance has gone too far."[43] Hartzog and Selinger suggest that

36 Heather Murphy, *She Was Going Into Labor. But She Had a Bar Exam to Finish*. The New York Times (September 13, 2020)(last accessed May 3, 2021)(also describing how students have resorted to wearing diapers or urinating in their seats).

37 Wong, *supra* note 33.

38 Brian Heaton, *State Legislatures Grapple with Biometrics Use in Schools*, Govtech Today (April 16, 2014), http://www.govtech.com/data/state-legislatures-grappling-with-biometrics-use-in-schools.html (last visited April 27, 2021).

39 LoSardo, *supra* note 34.

40 Barker, *supra* note 29.

41 Selwyn, *supra* note 1.

42 Tucker & Vance, *supra* note 34; *see also* LoSardo, *supra* note 34.

43 Tucker & Vance, *supra* note 34.

facial recognition invariably results in "impeding crucial opportunities for human flourishing."[44]

Additional concerns relate to the way that FRT and biometric technology make private information widely available in ways that were simply not possible in recent memory, eliminating the "practical obscurity" that used to exist when information was kept written down on sheets of paper and neatly stored in filing cabinets.[45] The way that, for example, faceprints and videos are compiled, structured, and stored in databases raises serious privacy and data protection concerns, not least because FRT often requires accessing material through a multimedia database.[46] While many systems will encrypt users' data and store it on their own data centers, on secured networks, or on the devices themselves, there is a genuine concern that data can be easily searched, repurposed and shared with third parties.[47] For example, it is possible that student data is unknowingly shared with third parties and then used as a virtual tracking device.[48] There is also a concern that public institutions like universities can be forced to turn over these data on public access grounds, especially in Nordic countries where the right to access public information is very strong.[49]

Data security is an urgent matter when it comes to online examination systems, which rely on highly sensitive biometric data as well as confidential data concerning exam material.[50] The reliance

44 Evan Selinger & Woodrow Hartzog, *The Inconsentability of Facial Surveillance*, 66 Loy. L. Rev. 33, 50 (2020).

45 Jonathan Turley, *Anonymity, Obscurity, and Technology: Reconsidering Privacy in the Age of Biometrics*, 100 B.U. L. Rev. 2179, 2247 (2020)("FRT and biometric technology nullify … practical obscurity with searchable databases. New private technology is rapidly eliminating anonymity even further.").

46 Elias Wright, *The Future of Facial Recognition Is Not Fully Known: Developing Privacy and Security Regulatory Mechanisms for Facial Recognition in the Retail Sector*, 29 Fordham Intell. Prop. Media & Ent. L.J. 611, 616–23 (2019).

47 Elizabeth A. Rowe, *Regulating Facial Recognition Technology in the Private Sector*, 24 Stan. Tech. L. Rev. 1, 24–34 (2020).

48 LoSardo, *supra* note 34.

49 Rowe, *supra* note 47.

50 Abdul Wahid, Yasushi Sengoku & Masahiro Mambo, *Toward Constructing a Secure Online Examination System,* IMCOM '15: Proceedings of the 9th International Conference on Ubiquitous Information Management and Communication (2015).

on public networks amplifies security concerns.[51] Biometric data can be the target of hacking or identity theft schemes.[52] When breaches occur in databases that contain large amounts of biometric data, the potential intrusion into the life of the individual is massive since it is extremely difficult to alter physiological characteristics.[53] Unlike a password or social security number which can be changed and replaced after a breach, there is almost no way to replace or remedy a breach involving biometric data.[54] It may be the case that stolen facial data can be used by a person with malicious intent to impersonate an individual.[55] Alarmingly, there have been notorious examples of breaches of large-scale biometric databases, for example, the fingerprint database maintained by the United States Office of Personnel Management.[56] Last year, there was also a data breach last that affected more than 440,000 individuals using the exam proctoring program ProctorU.[57]

Finally, it is important to mention concerns about the accuracy of proctoring systems that rely on the collection of biometric data. Unlike DNA or fingerprints, a person's face changes over time, and incorrect results can arise from the use of FRT: if a person gets a

51 *Id.*

52 Wright, *supra* note 46.

53 Wright, *supra* note 46.

54 Elizabeth McClellan, *Facial Recognition Technology: Balancing the Benefits and Concerns*, 15 J. Bus. & Tech. L. 363, 371–76 (2020); Fiona Q. Nguyen, *The Standard for Biometric Data Protection*, 7 J.L. & Cyber Warfare 61, 84 (2018); *see also*, Lauren Stewart, *Big Data Discrimination: Maintaining Protection of Individual Privacy Without Disincentivizing Businesses' Use of Biometric Data to Enhance Security*, 60 B.C. L. Rev. 349, 355–56 (2019) (explaining, "The value of biometric data lies in the data's unique and unchangeable nature, which provides much greater security than easily-hacked passwords.").

55 Wright, *supra* note 46; Daniel J. Solove & Danielle Keats Citron, *Risk and Anxiety: A Theory of Data-Breach Harms*, 96 Tex. L. Rev. 737, 757–58 (2018)(explaining,"Biometric data such as fingerprints or eye scans, health information, and genetic data cannot be exchanged. A criminal may obtain a victim's personal data and use it months or years later; the data will still be useful for committing fraud.").

56 Wright, *supra* note 46.

57 Brandon Paykamian, *Anti-Cheating Software Drawing Criticism at Universities*, Government Technology (April 09, 2021), http://www.govtech.com/education/higher-ed/anti-cheating-software-drawing-criticism-at-universities.html (last accessed April 27, 2021).

new hairstyle, grows some facial hair or gains some weight, then the technology may misidentify them.[58] In other words, false positives can arise when an individual makes even minor aesthetic changes.[59]

Non-discrimination

There is a substantial body of research that demonstrates that the use of FRT technologies threatens marginalized communities.[60] Study after study demonstrates that FRT is typically better at detecting light-skinned people than dark-skinned people, and better at detecting men than women.[61] This, of course, raises concerns that women or students of color will disproportionately and unfairly bear the consequences of these technologies.[62] Other groups at risk for discrimination by proctoring systems include: students with accessibility needs; students with learning disabilities, neurodivergence, and anxiety; low-income and rural students; and transgender students.[63]

58 Rowe, *supra* note 47; *see further*, Umar Toseeb, David R. T. Keeble & Eleanor J. Bryant, *The Significance of Hair for Face Recognition*, 7 PLoS ONE 1 (March 26, 2012), http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0034144 (last accessed April 27, 2021).

59 Marcus Smith, Monique Mann & Gregor Urbas, Biometrics, Crime & Security 64 (2018).

60 *See e.g.,* Joy Buolamwini & Timnit Gebru, *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*, Proceedings of Machine Learning Research (2018), http://proceedings.mlr.press/v81/buolamwini18a.html (last accessed April 27, 2021); *see also* Commission Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts, COM (2021) 206 final (Apr. 21, 2021) (hereinafter "AI Regulation").

61 Larry Hardesty, *Study Finds Gender and Skin-Type Bias in Commercial Artificial-Intelligence Systems*, MIT News (February 11, 2018), http://news.mit.edu/2018/study-finds-gender-skin-type-bias-artificial-intelligence-systems-0212 (last accessed April 27, 2021); Meredith Whittaker et al., *AI Now Report 2018*, AI NOW Institute, at 16 (December 2018) *citing* Buolamwini & Gebru, *id.*

62 Nila Bala, *The Danger of Facial Recognition in Our Children's Classrooms*, Duke L. & Tech. Rev. 249, 250–58 (2020).

63 Tyler Sonnemaker, *Tech Companies Promised Schools an Easy Way to Detect Cheaters During the Pandemic. Students Responded by Demanding Schools Stop Policing Them Like Criminals in the First Place*, Insider (November 1, 2020), http://www.businessinsider.com/proctorio-silencing-critics-fueling-student-protests-against-surveilalnce-edtech-schools-2020-10?r=US&IR=T (last accessed April 27, 2021).

Bias can arise both because of technical and social aspects. A core technical reason for why FRT technologies fail to identify people correctly is the use of training data sets that, for example, do not include people of African descent.[64] Selection bias is also an issue. For example, an FRT algorithm worked better on white people than on people of color because the images used for training the algorithm were collected by white developers.[65]

An additional problem is that AI is often seen as neutral and not subject to the biases of human beings.[66] Here, it is possible for an algorithm to be highly accurate yet be biased from a social point of view.[67] To put it differently, societal biases that exist in society can be reproduced in an algorithm.[68] From an ethical and legal point of view, it can be argued that AI should not just be "bias preserving" but also capable of improving the status quo.[69]

Current laws and regulations in Europe

This section will briefly explore existing and emerging laws that regulate the use of FRT in HE. As it currently stands, the legislative landscape in the EU associated with FRT in the HE context is highly complex and constantly evolving. Legal and ethical obligations are reflected in a number of binding legal instruments as well as in soft law and proposed legislation. There is no direct specific legal regime applicable to biometric data, other than the General Data Protection Regulation (GDPR). That said, there is a highly

64 Jay D. Aronson, *Computer Vision and Machine Learning for Human Rights Video Analysis: Case Studies, Possibilities, Concerns, and Limitations*, 43 Law & Soc. Inquiry 1188, 1194–95 (2018).

65 *Algorithms: Please Mind the Bias!* Institut Montaigne (March 2020), http://www.institutmontaigne.org/ressources/pdfs/publications/algorithms-please-mind-bias.pdf.

66 Bala, *supra* note 62.

67 Institut Montaigne, *supra* note 65.

68 Institut Montaigne, *supra* note 65.

69 Sandra Wachter, Brent Mittelstadt & Chris Russell, *Bias Preservation in Machine Learning: The Legality of Fairness Metrics Under EU Non-Discrimination Law*, W. Va. L. Rev. (forthcoming 2021), http://papers.ssrn.com/sol3/papers.cfm?abstract_id=3792772.

developed human-rights framework as well as an emerging regime which will govern high-risk AI like the use of FRT in the higher-education context.

Legal framework governing AI

Since the EU's working group on legal questions related to the development of AI and robotics was launched in 2015[70], the regulation of AI has become a hotly debated policy and academic subject. There have been intense discussions about whether AI needs specific regulation and, if so, what this regulation should look like. For example, some have argued that existing legal frameworks are sufficient to safeguard individuals from potential adverse effects of AI systems while others have contended that regulation is necessary but that it should take place at the Member State level instead of at the regional or international level.

A first step towards the regulation of AI occurred in April 2019 when the High-Level Expert Group on Artificial Intelligence (AI HLEG)[71] released its Ethics Guidelines on AI, setting forth four ethical imperatives in the context of AI: respect for human autonomy, prevention of harm, fairness and explicability.[72] From those four principles, seven principles to design trustworthy AI based on fundamental rights are derived which include: (1) human agency and oversight; (2) robustness and safety; (3) privacy and data governance; (4) transparency; (5) diversity, non-discrimination, and fairness; (6) societal and environmental well-being; and (7) accountability. Basically, in the guidelines the AI HLEG translated broad, normative assumptions and discussions into specific requirements which were reflected in soft law.

70 *See e.g.* Committee on Legal Affairs Working Group on Legal Questions related to the Development of Robotics and Artificial Intelligence, *Meeting of 22 October 2015 (Minutes)*, European Parliament, http://www.europarl.europa.eu/cmsdata/94927/Minutes_WG_Robotics_Oct.pdf.

71 The High-Level Expert Group on Artificial Intelligence (AI HLEG) was established as the flagship group for the European AI Alliance and tasked to provide guidelines on AI Ethics as well as AI policy and investment recommendations; for more *see* Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions: Artificial Intelligence for Europe, OJ 237 C 25.4.2018.

72 Nathalie A. Smuha, *The EU Approach to Ethics Guidelines for Trustworthy Artificial Intelligence*, Computer L. Rev. Int'l 97 (2019).

The next important step towards regulation occurred when the Commission released its White Paper on Artificial Intelligence. While the White Paper recognized the requirements set forth by the AI HLEG in its Ethics Guidelines, it expressly pointed to the need for regulation beyond soft law and provided guidance on how legislation should be developed to ensure legal certainty. More precisely, it proposed a risk-based regulation for AI with sector- and application-specific risk assessments and requirements as opposed to blanket requirements or bans.[73]

The White Paper explained that AI risks include risks for fundamental rights, including personal data and privacy protection and non-discrimination as well as risks for safety and the effective functioning of the liability regime.[74] It further stated that AI should be considered high risk when it is applied in a critical sector, and it is used in a manner in which significant risks are likely to arise.[75] Biometric identification was explicitly recognized as a high-risk application that should only be used "where such use is duly justified, proportionate and subject to adequate safeguards."[76]

In April 2021, the Commission put forward a legislative proposal maintaining the risk-based approach adopted in the White Paper, which broadly groups AI practices into four groups: unacceptable, high risk, limited risk and minimal risk. FRT is a central concern of the proposed AI regulation as it expressly prohibits the use of "real time" remote biometric identification systems in publicly accessible spaces for the purpose of law enforcement unless certain limited exceptions apply.[77] Outside of being used for law enforcement

73 *White Paper on Artificial Intelligence – A European Approach to Excellence and Trust* 16, European Commission (February 19, 2020), http://ec.europa.eu/info/sites/default/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf. (For high-risk AI applications the White Paper identifies six key requirements that could be included in upcoming AI legislation: (1) training data, (2) data and record-keeping, (3) information provision, (4) robustness and accuracy, (5) human oversight, and (6) specific requirements for certain particular AI applications, such as those used for purposes of remote biometric identification.).

74 *Id.*

75 *Id.*

76 *Id.* at 65, stating, "The gathering and use of biometric data for remote identification purposes, for instance through deployment of facial recognition in public places, carries specific risks for fundamental rights").

77 AI Regulation, *supra* note 60, Article 5(1)(d) and Explanatory Memorandum at 3.

purposes in publicly accessible spaces, Recital 33 and Annex III(1)(a) explain that "real-time" and "post" remote biometric identification systems should be classified as high-risk.[78] Consistent with the GDPR's definition, discussed in detail below, the regulation makes a sharp distinction between identification and verification techniques, placing stricter rules on the former[79], and essentially placing AI used for verification purposes outside the scope of high-risk AI all together.

Referring explicitly to the educational sector, Annex III states that AI systems used for "assessing students in educational training" constitutes high-risk AI.[80] It also refers to "AI systems intended to be used for the purpose of determining access or assigning natural persons to educational and vocational training institutions."[81] Recital (35) clarifies that the reason for this is that they "may determine the educational and professional course of a person's life and therefore affect their ability to secure their livelihood."[82] It is unclear whether Annex III's reference to "assessing students in educational training" refers to using AI to facilitate remote proctoring systems used to provide online assessments of students or whether it refers to using AI to literally assess – or score – students, through for example, some kind of grading software.

Where an AI system is deemed to be high-risk, then providers will have an extensive range of obligations.[83] Obligations for providers of AI systems include the adoption of risk management systems[84], data governance,[85] technical documentation[86], record-keep-

78 *Id.,* Recital 33.

79 *Id.*, Article 3(33); *see further* Recital 7 stating the definition in the AI Regulation should be interpreted consistently with Article 4(4) of the GDPR.

80 *Id.*, Annex III(3)(b).

81 *Id.,* Annex III(3)(a).

82 *Id.*, Recital 35.

83 See *Id.*, Chapter II.

84 *Id.*, Article 9.

85 *Id.*, Article 10.

86 *Id.*, Article 11.

ing[87], transparency[88], human oversight[89] and accuracy of outputs and security.[90] Additionally, there are a number of express obligations for providers of high-risk AI systems like putting in place a quality management system.[91] Many of these requirements must be performed *ex ante* before getting access to the EU market, which will ostensibly support a legal by design approach. Users of AI systems, like universities, also have explicit obligations like monitoring the operation of the high-risk AI system on the basis of the instructions of use.[92] Regulators will be able to fine non-compliant actors up to €30m, or 6% of their worldwide turnover.[93]

A timeline for the EU's AI Strategy

Legal framework for privacy and data protection

Article 8 of the European Charter of Human Rights (ECHR) sets forth a right to respect for private life. It covers four distinct areas: private life, family life, home, and correspondence.[94] Importantly, Article 8 imposes two types of obligations on the State. First, Article 8 obliges States to avoid interference with an individual's private life, family life, home and correspondence.[95] Second, there is a positive obligation to actively secure respect for private and family life, home and correspondence, between the state and the individual.[96] The positive obligation under Article 8 is derived from Article 1 ECHR,

87 *Id.*, Article 12.

88 *Id.*, Article 13.

89 *Id.*, Article 14.

90 *Id.*, Article 15.

91 *Id.*, Article 17.

92 *Id.*, Article 29(4).

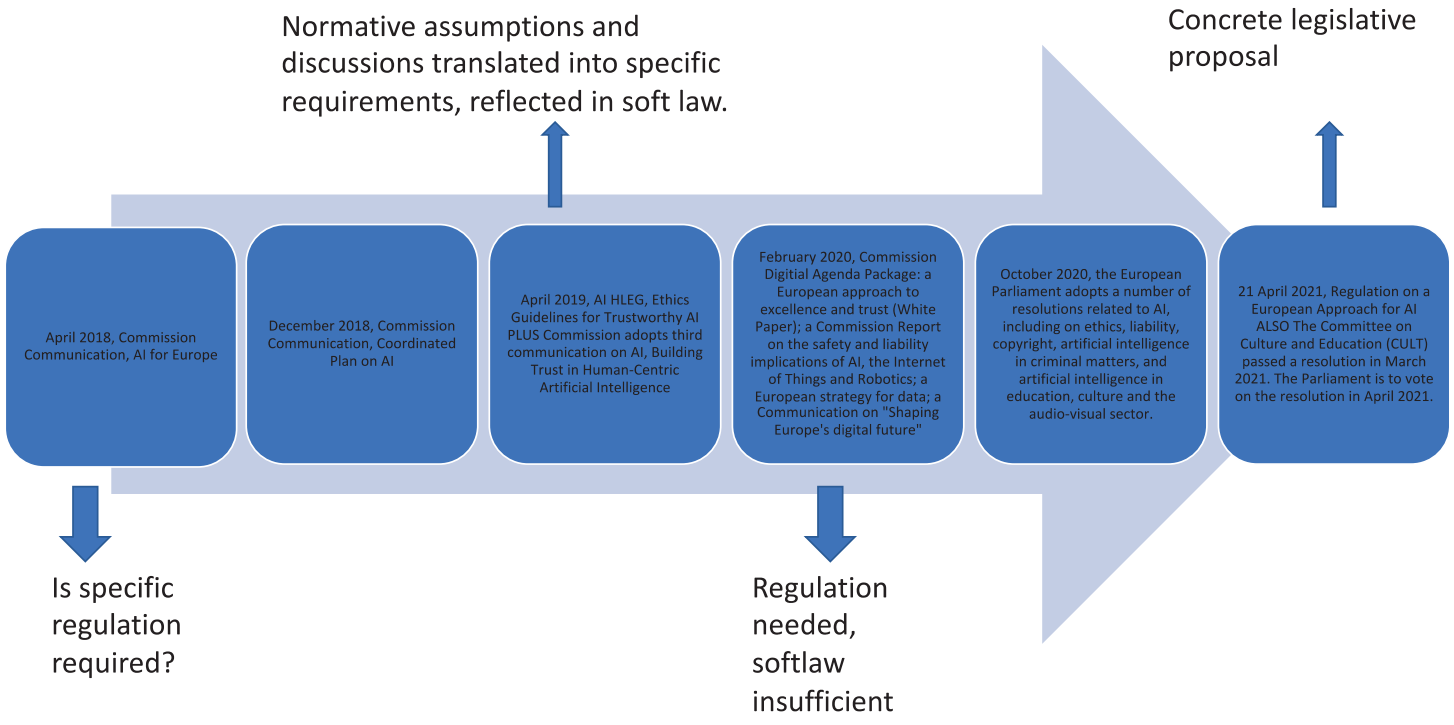
93 *Id.*, Article 71.

94 Ivana Roagna, *Protecting the Right to Respect For Private and Family Life Under the European Convention on Human Rights,* Council of Europe (2012), https://rm.coe.int/16806f1554.

95 *See e.g.* Leander v. Sweden, no. 9248/81, European Commission on Human Rights decision of 3 March 1987.

96 *See e.g.* I v. Finland, 2008 Eur. Ct. H.R., http://hudoc.echr.coe.int/eng?i=001-87510.

Current laws and regulations in Europe: AI



which requires states to secure Convention rights to everyone within their jurisdiction.[97]

In 2000, the EU proclaimed the Charter of Fundamental Rights of the European Union ("Charter") which became legally binding as EU primary law, pursuant to Article 6(1) of the TEU, when the Lisbon Treaty came into force on 1 December 2009.[98] Article 7 of the Charter reiterates the definition of privacy given by the ECHR.[99] Unlike the ECHR, however, the EU Charter defines the right to data protection as an autonomous right, instead of a simple dimension of the right to privacy.[100] It is important to emphasize that Article 8 not only explicitly mentions a right to data protection, but also refers to key data protection principles. It is further worth highlighting that the Charter requires that an independent authority will ensure compliance with the principles set forth in Article 8.

In 2016, the GDPR was adopted, modernizing EU data protection legislation and making it suitable for protecting fundamental rights in the digital age. The GDPR applies to partly or fully automatic AI systems that process personal data. It contains several core principles for the collection and processing of personal data such as lawfulness, fairness and transparency, purpose limitation, data minimization, accuracy, storage limitation, integrity and confidentiality (security) and accountability. It also affords the data subject numerous rights over their own personal data, such as the right to be informed of the collection and use of their personal data, the right to access their personal data and the right to have inaccurate or incomplete information corrected. As discussed in detail below, personal data must be processed lawfully in accordance with one of the six lawful grounds specified Article 6.[101]

97 Council of Europe (1952). European Convention for the Protection of Human Rights and Fundamental Freedoms, Europ.T.S. No. 5; 213 U.N.T.S. 221 (November 4, 1950), Article 1.

98 *See* consolidated versions of the European Communities (2012), Treaty on European Union, OJ 2012 C 326; and of European Communities (2012), TFEU, OJ 2012 C 326.

99 European Charter of Fundamental Rights, 2000 O.J. (C364), 18 December 2000, Article 7.

100 *Id.*, Article 8.

101 For all six legal bases see further European Union, Council Regulation 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of

Importantly, from the perspective of the use of proctoring in HE, Article 22 of the GDPR states that a data subject should not be subject to a decision based *solely* on automated processing, including profiling, which produces *legal effects* concerning him or her or *similarly significantly* affects him or her.[102] If an automated decision is likely to have a significant impact on the life of an individual, then special protection is necessary to avoid negative consequences. Automated decision-making includes profiling, which is defined in Article 4(4).[103]

Under the GDPR, a data controller has the responsibility to "implement appropriate technical and organizational measures", taking into account "the state of the art and the costs of implementation" and "the nature, scope, context, and purposes of the processing as well as the risk of varying likelihood and severity for the rights and freedoms of natural persons."[104] Article 32 explicitly refers to pseudonymizing and encrypting personal data as appropriate technical measures.[105] A legal definition of "pseudonymization" is set forth in the GDPR[106], which basically explains that pseudonymizing data means replacing the attributes in personal data – which make it possible to identify the data subject – with a pseudonym, and ensuring that the additional data necessary for reidentification are kept safely inaccessible for the users of "pseudonymized data."[107] This process can be juxtaposed with anonymization which requires all links to identifying the data subject to be broken.[108] Regularly testing

---

natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC, OJ 2016 L 119/1 [*hereinafter* General Data Protection Regulation], Art. 6.

102 *Id.* Article 22.

103 *Id.* Article 4(4).

104 *Id.* Article 32.

105 *Id.* Article 32(1).

106 *Id.* Article 4(5).

107 *Handbook on European Data Protection Law*, European Union Agency for Fundamental Rights 131, (2018); Waltraut Kotschy & Ludwig Boltzmann, *The New General Data Protection Regulation—Is There Sufficient Pay-Off for Taking the Trouble to Anonymize or Pseudonymize Data?* (November 2016), https://fpf.org/wp-content/uploads/2016/11/Kotschy-paper-on-pseudonymisation.pdf.

108 Article 29 Data Protection Working Party, *Opinion 05/2014 on Anonymization Techniques* (2014) WP 216, 3, 10 (http://www.pdpjournals.com/docs/88197.pdf (stating

and evaluating the effectiveness of the technical and organizational measures in place is also recommended.[109]

Legal framework for non-discrimination laws

While there is currently no AI specific equality legislation within European law, the EU has a well-developed *acquis communautaire*[110] of equality law. The EU has approved two major equality Directives[111] as well as adopted the Charter which includes anti-discrimination provisions set out in Chapter III. Furthermore, the Court of Justice (CJEU) has stated that equal treatment is a general or fundamental principle on which the EU is founded.[112] This body of law also draws on the jurisprudence of the European Court of Human Rights based on Article 14 of the ECHR.[113] Furthermore, by making a request for transparency pursuant to Article 15 of the GDPR, individuals may be able to identify that discrimination is occurring to the extent that a data controller must explain the categories of personal data being processed and the existence of automated decision-making, including profiling.

---

pseudonymization is "not a method of anonymization" but "merely reduces the linkability of a dataset with the original identity of a data subject, and is accordingly a useful security measure.").

109 GDPR, Article 32(1)

110 *Glossary of Summaries*, EUR-Lex, http://eur-lex.europa.eu/summary/glossary.html (last accessed April 27, 2021) (defining "acquis" as the "body of common rights and obligations that are binding on all EU countries, as EU Members.").

111 Council Directive 2000/43/EC implementing the principle of equal treatment between persons irrespective of racial or ethnic origin, OJ L 180, 19.7.2000, and Council Directive 2000/78/EC establishing a general framework for equal treatment in employment and occupation, OJ L 303 2.12.2000. Council Directive 2006/54/EC on the implementation of the principle of equal opportunities and equal treatment of men and women in matters of employment and occupation (recast), OJ L 204, 26.7.2006.

112 Case C-144/04, *Werner Mangold v. Rüdiger Helm* [2005] ECR I-09981; Case C-555/07, *Seda Kücükdeveci v. Swedex GmbH & Co. KG* [2010] ECR I-00365 at 20-22; and Case C-441/14 *Dansk Industri* (*DI*), *acting on behalf of Ajos A*/*S v Estate of Karsten Eigil Rasmussen,* OJ C 211 13.6.2016.

113 Robin Allen & Dee Masters, *Artificial Intelligence: The Right to Protection From Discrimination Caused by Algorithms, Machine Learning and Automated Decision-Making*, ERA Forum 585–598 (2020).

Legality case study

Personal data or sensitive data?

When personal data is gathered by a proctoring system it is subject to the GDPR, which defines personal data as "any information relating to an identified or identifiable natural person (data subject); an identifiable person is one who can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specific to his 5 physical, physiological, mental, economic, cultural or social identity."[114] To put it differently, if any information can be related to an identified or identifiable natural person then it is "personal." This is a broad definition and biometric data, like raw images of students, would clearly fall into this category, as they are inherently linked to a specific individual.

The EU also makes a distinction between personal data and sensitive personal data with the later receiving a higher level of protection under EU data protection. Sensitive data is information that relates to health, sex life, racial or ethnic origin, political opinions, religious or philosophical beliefs, and even trade-union membership. Under the GDPR, biometric data is explicitly covered under "special categories of personal data" and consequently, the processing of this data for the purpose of uniquely identifying a natural person is strictly forbidden, at least as a general principle.[115] That said, there are number of enumerated exceptions including, for example, obtaining explicit consent, specified public interest considerations, and certain exemptions in the fields of employment and social protection law.

The GDPR defines biometric data as "personal data resulting from specific technical processing relating to the physical, physiological or behavioral characteristics of a natural person which allow or confirm the unique identification of that natural person."[116] The reason for the general prohibition against processing biometric data is because biometrics, by their very nature, are "unlike other unique identifiers" since they are "biologically unique to the individual" and

114 GDPR, *supra* note 101, Article 4(1).

115 *Id.* Article 9(1).

116 *Id.* Article 4(14).

cannot be changed.[117] This means, as described above, if the data is compromised the risks of harm are very serious.

The classification of data used by proctoring exams as merely personal and subject to one of the Article 6 grounds for lawful processing or as sensitive and subject to one the Article 9(2) exceptions is of critical relevance. When understanding whether data falls under the definition of biometric data in the GDPR, it is first important to consider the source of biometric data. Here, it is important to understand that the GDPR protects two separate categories of biometric data.[118] First, it protects information connected to a person's physical or physiological trait like iris features or face patterns.[119] The second category concerns any behavioral information that can be used to uniquely identify someone, like the hand with which a person holds their phone.[120] Monajemi explains that it is unclear how the GDPR will regulate measures of a person's physical being based upon behavioral characteristics as "it has no nexus to the 'normal' definition of biometrics as it relates to body information."[121]

Second, in order to constitute sensitive data, the processing "needs to be carried out through specific technical means and measurements."[122] Recital 51 explains that the processing of digital photographs which may contain raw data relating to the physical characteristics of a person does not constitute biometric data unless it is "processed through a specific technical means allowing the unique identification or authentication of a natural person."[123] In other words, the image data might be used to create an individ-

117 Fiona Q. Nguyen, *The Standard for Biometric Data Protection*, 7 J.L. & Cyber Warfare 61, 84 (2018), citing 740 Ill. Comp. Stat. Ann. 14/5(c) (2008).

118 GDPR, *supra* note 101, Article 4(13).

119 *Id.* Article 4(1).

120 *Id.* Article 4(1); Michael Monajemi, *Privacy Regulation in the Age of Biometrics That Deal with A New World Order of Information*, 25 U. Miami Int'l & Comp. L. Rev. 371, 383 (2018).

121 *Id.*

122 GDPR, *supra* note 101, Article 4(14).

123 GDPR, *supra* note 101, Recital 51 (stating, "The processing of photographs should not systematically be considered to be processing of special categories of personal data as they are covered by the definition of biometric data only when processed through a specific technical means allowing the unique identification or authentication of a natural person…").

ual digital template or profile, which in turn is used for automated image matching and identification.[124] The key to a finding of biometric data is the existence of biometric processing, and not just the existence of a database with facial images or fingerprints.[125] Kindt explains: "Facial images only become biometric data … if they are used for biometric comparison, and more precisely, if they are the result of 'specific technical processing'."[126]

Third, the purpose of the processing must be identified insofar as the GDPR distinguishes between processing biometric data for identification purposes and verification purposes. Here, Article 9(1) GDPR only includes "biometric data for the purpose of uniquely *identifying* a natural person." In other words, if biometric data is processed for the purpose of verification, which does not aim to uniquely identify a natural person, the processing would not fall within the prohibition provided for in Article 9(1) GDPR.

Biometric identification can be described as "using an individual's biometric identifier to match the identifier with that specific individual within a database of biometric identifiers compiled from multiple individuals."[127] Essentially, it seeks to answer the question: who is this student? In the education context, this process would permit the unique identification of a student from a database containing data on all students in a given population, confirming his or

124 *What is Special Category Data?*, Information Commissioner's Office, http://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/special-category-data/what-is-special-category-data/#scd4 (last accessed April 27, 2021).

125 Els J. Kindt, *Having Yes, Using No? About the New Legal Regime for Biometric Data*, 34 Computer L. & Security Rev. 523 (2018).

126 *Id.* (providing the example, "Photographs, for example of children at schools, if collected and disclosed on websites of the school, or registered in an internal database of the school, are according to the new definition in the GDPR not biometric data as long as they are not processed by a biometric system.").

127 Kelly A. Wong, *The Face-Id Revolution: The Balance Between Pro-Market and Pro-Consumer Biometric Privacy Regulation*, 20 J. High Tech. L. 229, 232 (2020); *see also* J. Valera, J. Valera and Y. Gelogo, *A Review on Facial Recognition for Online Learning Authentication*, 8th International Conference on Bio-Science and Bio-Technology 16-19 (2015) ("…user identification determines the person based on exhaustive verification where the actual biometric features is compared to all registered references and determined of which has the greatest similarity.").

her identity (positive identification).[128] It might also identify that a student's information is not present in a certain database, preventing him or her from having multiple identities in the system (negative identification).[129] Sometimes this approach is referred to as a one-to-n matching process, where "n" is the total number of biometrics in the database.[130] Ankerman explains: "Both positive and negative identification serve the same goal: to authenticate each individual based on a single, non-transferable identity."[131] Importantly from a data-protection standpoint, for identification to function, it is always necessary to utilize a database of stored biometric data, as compared to just the storage of a single biometric characteristic.[132]

On the other hand, in biometric verification, an individual's biometric trait is scanned and compared to the existing template that has been formed for that specific individual to verify that the individual is who he or she claims to be.[133] Essentially, it seeks to answer the question: "Is the student who they claim to be?" Sometimes this approach is referred to as a one-to-one matching process.[134] With biometric verification, it is only necessary to store a single biometric characteristic.[135] This data may be stored in a database or stored locally, on an identification card, for example.[136] The Council of Europe has explained that biometric verification contains less risk

128 Stefan P. Schropp, *Biometric Data Collection and Rfid Tracking in Schools: A Reasoned Approach to Reasonable Expectations of Privacy*, 94 N.C. L. Rev. 1068, 1071–72 (2016); James Wayman et al., Biometric Systems 5 (2005).

129 Alexa N. Acquista, *Biometrics Takes Off – Fight Between Privacy and Aviation Security Wages On*, 85 J. Air L. & Com. 475, 480 (2020).; Wayman, *id.*

130 Kindt, *supra* note 125; Margaret Hu, *Biometric Id Cybersurveillance*, 88 Ind. L.J. 1475, 1491 (2013).

131 Chantelle D. Ankerman, *A Closer Look: Iris Recognition, Forensics, and the Future of Privacy*, 49 Conn. L. Rev. 1357, 1379 (2017).

132 Acquista, *supra* note 129.

133 Clifford S. Fishman & Anne T. McKenna, Biometrics: A General Overview of Biometric Technology in Wiretapping and Eavesdropping (2019).

134 Kindt, *supra* note 125 ("The processing for verification purposes is a one-to-one (1:1) comparison and is used to verify and to confirm by biometric comparison whether an individual is the same personas the one from whom the biometric data originates.").

135 Els J. Kindt, Privacy and Data Protection Issues of Biometric Applications: A Comparative Legal Analysis 18, 38-39 (2013).

136 *Id.*

than biometric identification because the utilization of a database is not required.[137]

In the context of proctoring exams, it may be possible that emotional data like the facial expressions of students (without the retention of facial image characteristics) would constitute mere personal data if it is insufficiently distinctive to allow or confirm identification of the student.[138] Furthermore, behavioral data that is insufficiently distinctive to allow or confirm identification would also fall in this category.[139] Personal data that relates to a student's physical, physiological or behavioral characteristics, which allows for unique identification or confirmation of an identity, but are not used in a biometric comparison, also fall within the scope of personal data.[140] Furthermore, biometric data like iris features and face patterns captured by a proctoring system will not constitute biometric data if they are used in biometric verification systems whereby only a one-to-one comparison is made based on biometric data. However, using biometric data in a proctoring system that involves identification in a one-to-many matching process is in principle forbidden, unless exempted.

Finally, the intersectionality that exists between facial image data and special categories of data must be highlighted. That is, data collected by proctoring systems, particularly, facial image data, may reveal information about racial or ethnic origin, religious orientation, health related information and even sexual orientation. This information may be shared, consciously or unconsciously. The legal significance of this that certain data collected by these systems may nevertheless be classified as sensitive data even through it falls outside the technical definition of "biometric data" in the GDPR.[141]

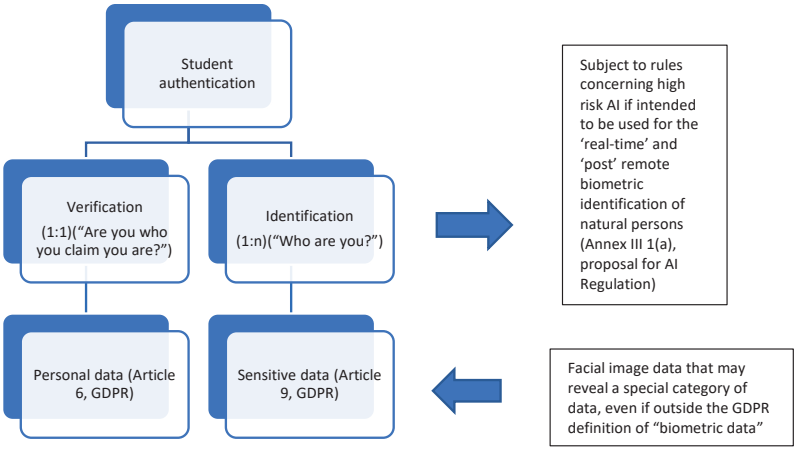
137 *See Progress Report on the Application of the Principles of Convention 108 to the Collection and Processing of Biometric Data*, Council of Europe (2005), and the updated Progress Report of 2013, T-PD(2013)06, https://rm.coe.int/progress-report-on-the-application-of-the-principlesof-convention-108/1680744d81 (last accessed April 27, 2021); *see also* Kindt, *supra* note 125 (explaining, "The distinction between these two functionalities, whereby identification requires a data-base with one or more data records, is of key importance in the discussion and regulation of biometric data processing.")

138 Kindt, *supra* note 125.

139 Kindt, *supra* note 125.

140 Kindt, *supra* note 125.

141 Kindt, *supra* note 125.



Lawful grounds

Article 6, personal data

As explained above, processing personal data requires a lawful basis. The lawful bases for processing personal data in the context of online proctoring are namely: (1) consent, (2) the need to process personal data in order to perform a task that is in the public interest or under public authority, and (3) the need to process personal data in the context of a legitimate interest.

Consent

Relying on consent as a valid ground to process personal data in connection with online proctoring is generally not possible because of the power imbalances and the hierarchical relationship that exists between the students and the teachers representing the university.[142] Students might feel coerced to give consent because they fear they will get a bad grade. This is particularly true when they are under exam pressure, exacerbated by a pandemic as well as the surveillance capabilities of the exam software itself. Furthermore, in a lock-down situation, it is not clear that students can freely offer consent to

142 *See* GDPR, *supra* note 101, Article 4(11)(defining consent).

the extent that no practical alternatives may be available for taking tests.[143]

While the requirements for consent to be freely given, informed, specific and unambiguous are difficult to meet in the university context, it may be used in certain, limited situations, for example, for students who wish to take exams from abroad.[144] It is also possible that the consent grounds to process personal data takes on a bigger role when alternative legal grounds like legitimate interest and public interest, discussed more below, are not strictly required by the exigencies of the pandemic situation. In the post-pandemic context, for example, the consent grounds might be used for students who have indicated that they prefer to take their exams at home, because of noise that exists in an exam hall or because they want to skip the commute to the university.[145] If consent is relied upon then it is important to remember that students may revoke their consent at any time, and the university has an obligation to keep a record of the consent.

Legitimate interest

When considering the interests of online proctoring, it is first required to consider the legitimate interest of the institution. Second, the extent to which the processing is necessary to defend the legitimate interest must be analyzed with references to the concepts of proportionality and subsidiarity. Basically, a balance between the institutions interests and the students' rights to privacy and data protection must be struck which must be well documented.

According to the GDPR, the legitimate interest basis may not apply to processing carried out by public authorities as part of their tasks.[146] As such, an institution that classifies itself as a government organization does not have the option to use legitimate interest as a basis. That said, in the context of the coronavirus crisis, VU Amster-

143 *Student Proctoring Software Gets First Test Under EU Privacy Law*, Bloomberg Law (July 29, 2020), http://news.bloomberglaw.com/tech-and-telecom-law/student-proctoring-software-gets-first-test-under-eu-privacy-law (last accessed April 27, 2021).

144 *Whitepaper Online Proctoring: Questions and Answers At Remote Surveillance*, SURF (April 2020), http://www.surf.nl/files/2020-06/surf-whitepaper-online-proctoring_en_mei-2020.pdf.

145 *Id.*

146 GDPR, *supra* note 101, Article 6 ("Point (f) of the first subparagraph shall not apply to processing carried out by public authorities in the performance of their tasks.").

dam first argued that it is not a government agency due to private funding, and second, asserted that it had a legitimate interest to process personal data for online proctoring.[147] Furthermore, even some clearly public universities have argued that a public authority performing activities that are not part of a public task may do so on the basis of legitimate interests.[148]

Legitimate interests include things like the need "to organize exams and avoid postponement of exams as much as possible, as this leads to study delay for students, accumulation of work for teachers and a shortage of spaces for taking exams at a later stage" as well as needed to fulfil the requirements of distance education by securely organizing remote exams.[149] However, online proctoring will likely be found to be disproportionate after universities open and exams can be held in halls again. This is because a less privacy intrusive alternative to online proctoring would exist.[150]

Task of public interest

A university may invoke as a legal basis the argument that processing is necessary for the performance of a task carried out in the public interest or in the exercise of official authority vested in the data controller.[151] The GDPR requires that the underlying task, function or power of the public body must have a clear basis in law. The "law" in question, however, does not have be a legislative act adopted by

147 *Guide to Reference Framework For Alternative Forms of Assessment*, Vrije University Amsterdam (March 26, 2020)("B. The use of online surveillance by means of video images VU Amsterdam can use online surveillance without the student's permission, provided that: 1. a case can be made that remote monitoring of the assessment in question is needed to: check the identity of the individual taking the assessment; establish that no academic misconduct has been committed during the assessment; and to establish that the assessment was completed within the allotted time frame… In such cases, VU Amsterdam can be said to have a necessary, legitimate interest that outweighs the rights and freedoms of those involved (Art. 6.1 f GDPR).")

148 Rechtbank Amsterdam 11 June 2020 rolnr. C/13/684665; *Court Decision on Remote Proctoring in the Netherlands*, Association of Test Publishers, https://www.testpublishers.org/amsterdam-court-case (last accessed April 27, 2021).

149 Meike Davids, *Data Protection Impact Assessment (DPIA),* U. of Twente (December 17, 2020), http://www.utwente.nl/remote-exams/students/proctoring/dpia-proctoring.pdf.

150 SURF, *supra* note 144.

151 GDPR, *supra* note 101, Article 6(1)(f).

parliament but can, for example, include a university charter.[152] Therefore, things like advancing education, learning and research to be a public task can be considered part of the "public tasks" of a university. Here, there can be little doubt that universities have legal obligations to administer exams, award degrees and make efforts to prevent fraud/ensure the quality of the education in doing so.[153]

The principles of proportionality and subsidiarity are key issues when it comes to online proctoring exams. First, the processing of personal data must be necessary and in proportion to the ends (proportionality). Second, if the university could reasonably perform its tasks or exercise its powers in a less intrusive way, this lawful basis does not apply. As discussed above, online proctoring has the potential to intrude deeply into the private lives of students through, for example, AI that monitors the student's computer, home, and facial images. The essential question becomes where to draw the line, and whether the use of FRT in the educational context in order to ensure academic integrity and to verify the student's identity are necessary and proportionate.

In a pre-Covid 19 case, the Danish Data Protection Agency found that a high school's reliance on an online proctoring system (Examcookie), did not sufficiently explain that the processing of the collected information about all examinees had been sufficient, relevant and limited to what is necessary in relation to the purpose of detecting and preventing fraud.[154] It found that the high school, Fredericia Gymnasium, only explained that there was a need to prevent exam cheating, but that the monitoring of the students' private computers through the proctoring system intruded too deeply into the private

152 *Id.* Recital 41 (clarifying that a "legal basis" does not have to be an explicit statutory provision, as long as the application of the law is clear, precise and foreseeable.).

153 Selwyn, *supra* note 1; see for an example, Rechtbank Amsterdam, *supra* note 148 ("In this case, the public task of the UvA is regulated by, or can be traced back to, a statutory task, namely its task to provide education, to conduct exams and to issue diplomas, while maintaining the quality of that education and of the diplomas is guaranteed. This task is detailed in the WHW, and the authority to process data in the context of exams is further detailed in the OER and in the Rules and Guidelines of the Examination Board.").

154 *Fredericia Gymnasiums behandling af personoplysninger ved brug af programmet Examcookie*, Datatilsynet (May 16, 2019), https://www.datatilsynet.dk/afgoerelser/afgoerelser/2019/maj/fredericia-gymnasiums-behandling-af-personoplysninger-ved-brug-af-programmet-examcookie (last accessed April 27, 2021).

lives of the students.[155] The Danish Data Protection Agency emphasized that Fredericia Gymnasium did not account for all the personal information collected being necessary to fulfill Fredericia Gymnasium's purpose of detecting and preventing fraud.[156]

In another more recent case, however, the Danish Data Protection Agency found that the use of an online proctoring system (ProctorExam) to control cheating during exams was necessary and proportionate.[157] The Danish DPA emphasized the COVID-19 situation and the fact that the IT University (ITU) was physically closed and forced to conduct all teaching and all examinations online.[158] It also emphasized that ITU reportedly made an assessment of the need for examination supervision for different subject areas and found that in the subject in question ("Algorithms and Data Structures"), it was particularly necessary to utilize online proctoring.[159] In other words, ITU only used ProctorExam for exams where it was specifically deemed necessary. This case demonstrates that what is necessary for the performance of a public task may vary over time to the extent that a university may come to the conclusion that it is required to process video images of students in specific situations in order to carry out the public tasks laid down by their charters during a pandemic, unlike in normal times.

Article 9, Sensitive data

If data collected from a proctoring system is classified as biometric data, then it should not be processed unless one of the ten exemptions from the prohibition of processing biometric data found in Article 9(2) apply. This section will explore how these exemptions might apply in the context of proctoring exams. The first exemp-

155 *Id.*

156 *Id.*

157 *Universitets brug af tilsynsprogram ved online eksamen*, Datatilsynet (January 26, 2021), https://www.datatilsynet.dk/afgoerelser/afgoerelser/2021/jan/universitets-brug-af-tilsynsprogram-ved-online-eksamen (last accessed April 27, 2021).

158 *Id.*

159 *Id.* (reasoning that "the subject was a basic course where the students had to show their basic skills within the subject area. All correct answers in the exam would be identical, as there was one correct answer without explanation or elaboration, which is why, unlike other subjects, it was crucial to be able to demonstrate that the examinee did not receive help from others.").

tion concerns where the "explicit consent" of the data subject is obtained.[160] For the reasons set forth above, this exemption will, at best, have narrow application. This is particularly true in light of a 2019 decision by the Swedish Data Protection Authority, which rejected consent as a valid ground for a school in Northern Sweden to use FRT to keep track of students' attendance in school based on the relationship of dependence between students and institutions, as well as the substantial power imbalance between the different actors.[161] It is also important to note that Union or Member State law may limit the circumstances where explicit consent can be relied upon as a legal grounds to process biometric data.[162] Kindt notes: "This leaves Member States to carefully think about situations where biometric identification, based on consent, may not be desirable."[163]

Another exemption is found in Article 9(2)(g) where the processing is necessary "for reasons of substantial public interest" so long as there is a Union or Member State law which is (1) proportionate to the aim pursued, (2) respects the essence of the right to data protection and (3) provides for suitable and specific measures to safeguard the fundamental rights and the interests of the data subject.[164] When discussing whether this ground applied in the Fredericia Gymnasium case, the Danish DPA rejected that the school had a ground for processing sensitive personal data covered by Article 9 of the GDPR.[165] However, in the ITU case, the DPA found that Article 9(1)(g) was a legal basis for processing sensitive data. The DPA emphasized that the ITU did not rely on image identification at the beginning of the examination. More specifically, it stated "The ITU does not use software-based face recognition or other technical treatment to uniquely

160 GDPR, *supra* note 101, Article 9(2)(a).

161 *Supervision Pursuant to the General Data Protection Regulation (EU) 2016/679 – Facial Recognition Used to Monitor the Attendance of Students, Ref. no. DI-2019-2221*, Swedish Data Protection Authority (August 20, 2019), http://www.imy.se/globalassets/dokument/beslut/facial-recognition-used-to-monitor-the-attendance-of-students.pdf.

162 GDPR, *supra* note 101, Article 9(2)(a).

163 Kindt, *supra* note 125.

164 GDPR, *supra* note 101, Article 9(2)(g).

165 Datatilsynet, *supra* note 154.

identify the examinee."[166] Instead, at random checks, staff from the ITU manually checked the identity of the examinee by holding a student card or other photo ID up to the face, which appears on the computer camera.[167] It also emphasized that the ITU "encouraged the examinees to arrange their computers in such a way as not to accidentally process sensitive information in connection with the recordings of video, audio and screen."[168]

If a proctoring system involves the processing of a special category of personal data on a large scale, then a DPIA will be required.[169] Prior consultation with the supervisory authority will also be required in the event that the DPIA indicates that the processing would result in a high risk in the absence of or which the controller cannot mitigate by appropriate measures 'in terms of available technology and costs of implementation'.[170] Consultation with the supervisory authority is also required if national law requires prior authorization for a task carried out in the public interest.[171]

Relevant factors for determining whether the use of AI-based proctoring tools are necessary and proportionate to achieving their aims

The question of which legal basis is appropriate in a specific situation when utilizing online proctoring will always depend on the circumstances, the concrete purpose for the use of the tool, and the type of data being processed. It is of utmost importance that the educational institution is able to justify its choice of a particular legal basis. In addition, the processing of personal data must be necessary and proportionate to achieve the underlying purpose.

166 Datatilsynet, *supra* note 157.

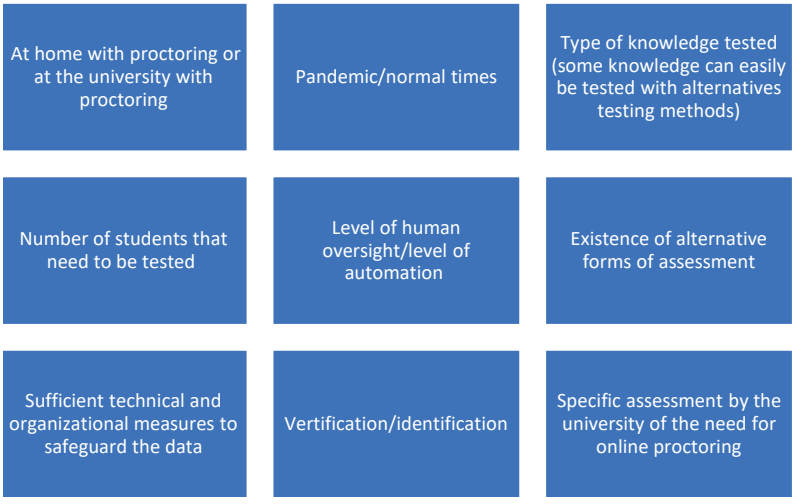
167 *Id.*

168 *Id.*

169 GDPR, *supra* note 101, Article 35.

170 *Id.* Article 36.

171 *Id.* Article 36(5).



Conclusion

In light of the potential harms posed by online proctoring exams, one potential response is to reject them completely. Several prominent groups and privacy experts have made it clear that educational institutions "must proceed with great caution when considering the implementation of FRT, especially to provide safeguards necessary to protect students against its many projected, yet ultimately unknown harms."[172] For example, Hartzog and Selinger call for a wholesale ban on facial recognition, explaining that "when technologies become so dangerous, and the harm-to-benefit ratio becomes so imbalanced, categorical bans are worth considering."[173] Likewise, Barrett calls for ban of FRT in the school context because of the "vast and far-reaching" harms.[174] Complete bans already exist in some American schools.[175]

172 LoSardo, *supra* note 34.

173 Woodrow Hartzeg & Evan Selinger, *Facial Recognition is the Perfect Tool for Oppression*, MEDIUM (August 2, 2018), http://medium.com/s/story/facial-recognition-is-the-perfect-tool-for-oppression-bc2a08f0fe66 (last accessed April 27, 2021).

174 Lindsey Barrett, *Ban Facial Recognition Technologies for Children-and for Everyone Else*, 26 B.U. J. Sci. & Tech. L. 223, 275–83 (2020).

175 *See* Blake Montgomery, *Facial Recognition Bans: Coming Soon to a City Near You*,

In the EU, many are pushing for a multi-year moratorium on FRT so that the technology's impact can be studied.[176] While it appears that this idea was at least explored in earlier drafts of the AI White Paper[177], no references to a ban or a moratorium were reflected in the final published document nor have any been included in the AI Regulation. Instead, the EU appears to be adopting a risk-based approach regulation to these types of technologies.

Ultimately, future scenarios of online proctoring in a post pandemic world remain unclear. Will it be immediately cut by the institutions themselves in order to implement a clear and consistent policy, in favor of the protection of students' human rights? Will it fade out as regulators make clear that legal grounds to process personal data in these systems do not exist? Will it be continued to be used because of increased demand for digital education, as well as for its potential to cut costs like the maintenance of physical computers, physical exam space and live proctors?

The Daily Beast (July 31, 2019), http://www.thedailybeast.com/facial-recognition-bans-coming-soon-to-a-city-near-you (last accessed April 27, 2021).

176 *See* Janosch Delcker, *Activists Urge EU to Ban Live Facial Recognition in Public Spaces*, Politico (November 12, 2020), http://www.politico.eu/article/activists-urge-eu-to-ban-live-facial-recognition-in-public-spaces/ (last accessed April 27, 2021).

177 *Facial Recognition: EU Considers Ban of Up to Five Years*, BBC News (January 17, 2020), http://www.bbc.com/news/technology-51148501 (last accessed April 27, 2021).

A Conceptual Approach to AI and Data Protection

CECILIA MAGNUSSON SJÖBERG

1 Introduction

In this brief contribution to the Nordic Yearbook in Law and Informatics, I will address what may be referred to as a conceptual approach to artificial intelligence (AI) and data protection.[1] It is of course not possible to here deliver a complete solution to the challenges related thereto; the goal is rather to shed light on topical discussions. Another way to express this ambition is in terms of

1 Parts of this text were presented at the XXXV Nordic Conference on Law and IT, 11–12 November 2020, e-Stockholm 2020. Some other parts were presented at the CPDP (Computers Privacy and Data Protection) Conference on 27 January 2021, Brussels.

getting a grip on the General Data Protection Regulation (GDPR)[2] and other data protection legislation, taking AI into consideration. So, this is not a work of jurisprudence but rather of *legal informatics* within the broader framework of law and information and communications technology (ICT).[3]

In this context, a conceptual approach leads the way towards *legal system management,* supplementing a traditional dogmatic legal analysis based on sequential rules and regulations (at different levels of a norm hierarchy), as well as settled court cases. The hypothesis here is that a broader *interpretative scope* facilitates for both beginners and advanced legal scholars and practitioners to apply law in a better way than would otherwise be possible.[4] The impact of AI is substantial from a conceptual point of view, not least when it comes to big data,[5] machine learning (ML), natural language processing (NLP), etc.

Returning to aforementioned conceptual approach, *transparency* is no doubt a condition for *privacy* in the context of *personal data processing* based on *AI* methods[6] (see Article 5.1 (a) GDPR, Princi-

2 Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (hereinafter 'GDPR').

3 There are many ways to capture the interplay between law and ICT. Here are two references:

Kevin D. Ashley, Artificial Intelligence and Legal Analytics: New Tools for Law Practice in the Digital Age (2017), and Andrew Murray, Information Technology Law: The Law & Society (2019).

In general terms, the first reference is oriented towards methodological issues and the second one focuses more on substantive law.

4 There is a labyrinth of labels to navigate if the conceptual approach is found attractive. To mention but a few: conceptual model, decision tree, ontology, taxonomy, terminology, and vocabulary. For further illumination on this, the reader is recommended literature such as Cecilia Magnusson Sjöberg, Critical Factors in Legal Document Management: A Study of Standardised Markup Languages (1998).

5 See further Liane Colonna, Legal Implications of Data Mining: Assessing the European Union's Data Protection Principles in Light of the United States Government's National Intelligence Data Mining Practices (2016).

6 Information on recent developments can be found in the Commission Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts, COM (2021) 206 final (Apr. 21, 2021) (hereinafter 'AI Regulation', sometimes called 'Artificial Intelligence Act'). See also Cecilia Magnusson Sjöberg, *Legal Automation: AI in Law Revisited*, in Legal Tech, Smart Contracts

ples relating to processing of personal data). However, another point should be made regarding data protection principles: Processing governed by *principles of openness* might not provide transparency due to a lack of *access rights* and how they are (insufficiently) implemented (see Article 25 GDPR, Data protection by design and by default).[7]

The structure of this contribution to the yearbook (1), the methodological approach to clusters of concepts will be presented (2). Attention is then shifted to acronyms as a lever for efficiency (3). After this introduction the reader will be introduced to the interplay between AI and data protection by design and default (DPbDD) (4), and to a Nordic law perspective (5). Emphasis will be placed on transparency by way of so-called routine measures (5.1), records management (5.2), and technical interpreters (5.3). Lastly, there will be some concluding remarks (6) on the DataLEASH project (6.1), as well as forthcoming research (6.2). References are listed in footnotes, providing information on the emerging landmarks of law and AI.

2 Clusters of concepts

In this investigation of the potential of a conceptual approach to legal interpretation and application, it should be mentioned that the study objects are not always separate concepts, but rather *clusters* of them.[8] The substantive domain is still (personal) data protection – in particular, but not exclusively, with reference to Article 4 GDPR – with some connections to AI. The enumeration below is of course not exhaustive, but serves the purpose of conceptual exemplification.

---

and Blockchain pp. 173–187 (Marcelo Corrales, Mark Fenwick & Helena Haapio eds., 2019). For a broader perspective see, for instance, EU och teknologiskiftet: Europaperspektiv 2020 (Antonina Bakardjieva Engelbrekt, Anna Michalski & Lars Oxelheim eds., 2020). Current critical discussions are also to be found in Kate Crawford, Atlas of AI. Power, Politics, and the Planetary Costs of Artificial Intelligence (2021).

7 See also Cecilia Magnusson Sjöberg, *Legal AI from a Privacy Point of View: Data Protection and Transparency in Focus*, in Digital Human Sciences (Sonya Petersson ed., 2021).

8 See also Core Concepts and Contemporary Issues in Privacy (Ann E. Cudd & Mark C. Navin eds., 2018).

- **Formal**

There are many concepts that are relevant to the data protection domain. Some are formal, others are stipulative, i.e., made up in a certain situation. Formal *legal definitions* exist for 'personal data', 'controller', and 'consent', while e.g., 'synthetic data' are extensively discussed, but not formally defined.

- **Functional**

An example of a functional concept is 'processing'. The underlying legal definition is broad, but a more in-depth analysis shows that there are quite specific *legal consequences* to be aware of. For instance, 'anonymisation' is one way to avoid being included in the scope of the GDPR, while 'pseudonymisation' and 'deidentification' (and, also, 'reidentification') are all privacy-enhancing measures that take place within the material scope of this EU Regulation.

- **Steering**

Without being (logically) represented in the form of a rule, some concepts indicate the use of certain procedures. *Algorithms* – and associated models of different kinds – are one example hereof.[9] An 'algorithmic' approach may be described as a structured way of problem-solving with software, entailing code[10] that in combination with (big) data can be run by computers. To put it simply: algorithms can be static and deterministic or dynamic and self-learning. Algorithms may, with reference to the ongoing work with the EU proposal concerning AI (see above), be based on so-called training data (Commission's proposal Article 29), validation data (Commission's proposal Article 30) or testing data (Commission's proposal Article 31).[11]

9 On algorithms in the legal domain, see the decision of the Swedish Ombudsman JO, Dnr 6783-2019, Beslut 2021-06-09. *Kritik mot Arbetsmarknadsnämnden i Trelleborgs kommun för dröjsmål med att lämna ut en skärmdump* (Swedish).

10 Read more about the OECD's work to promote rules as code (RaC) in the public sector: James Mohun & Alex Roberts, *Cracking the Code: Rulemaking For Humans and Machines*, OECD Working Papers on Public Governance No. 42, OECD (2020), available at https://www.oecd-ilibrary.org/governance/cracking-the-code_3afe6ba5-en (last accessed July 6, 2021).

11 A common understanding has been that ICT can merely provide tools, like a calculator. As digitalisation develops, the understanding of ICT is increasing in society. A sign of this is a Swedish public (governmental) inquiry proposing to amend the Swedish Public Access and Secrecy Act (2009:400) so as to explicitly regulate the use of algorithms in the public sector: Section 3 a 'En myndighet ska se till att informa-

- **Topical**

Some concepts are topical. Legal aspects of *information security* are a good example of this, comprising core concepts like 'confidentiality', 'integrity' and 'availability'. Mention should also be made of 'traceability', 'non-repudiation', 'accountability', etc. Confidentiality has a technical meaning, as well as a legal one denoting contractual and regulatory requirements. Accomplishing integrity is not necessarily related to privacy (a right to be left alone), but rather to bringing about a high level of data quality with regard to information security, etc. At a generic level, this calls attention to the need for a linguistic interface between law and technology.

- **Historical**

Furthermore, it can be enlightening to add a limited historical perspective to a discussion on data protection. Given the fact that the history of law-making over the years has provided us with quite a few concepts, some understanding of legacy is justified, to avoid unnecessary problems related to either ICT or regulation. For instance, Sweden was the first country in the world to successfully enact a national Data Protection Act (SFS 1973:289). In the beginning, emphasis was entirely on personal data files. In response to the now repealed EU Data Protection Directive (95/46/EC), Sweden was also the first to introduce a so-called *misuse model* (Section 5 a, SFS 1998:204), with an easy track for compliance when there was no risk of privacy infringements from unstructured personal data. However, the introduction of the GDPR showed that Sweden's pragmatic approach to this matter was in vain. No such easy roadmap for compliance can be found in the current EU Regulation, regardless of how trivial and harmless the personal data processing might be. Each provision must be complied with, by both the controller and the processor.

---

tion kan lämnas om hur myndigheten vid handläggning av mål eller ärenden använder algoritmer eller datorprogram som, helt eller delvis, påverkar utfallet eller beslutet vid automatiserade urval eller beslut' [An authority shall ensure that information can be provided on how the authority, in handling cases or errands, uses algorithms or computer programs that, fully or partially, affect the outcome of automated selections or decisions], p. 34 SOU 2018:25, *Juridik som stöd för förvaltningens digitalisering. Betänkande av Digitaliseringsrättsutredningen* (Swedish).

3 Acronyms as a lever for efficiency

At this stage, a few lines on writing style within the subject area are justified. More precisely, this concerns the use of acronyms as a supplementary tool for communication by way of full sentences. This might sound silly, but for those concerned, there are both possibilities and pitfalls to shortening text in this way. If you are in command of the glossary applied in a case, whatever the legal issue and activity, rapid and insightful writing and reading will be the outcome, which are worthwhile goals. This also shows that you are professionally qualified. Otherwise, there is a risk of enhanced and time-consuming uncertainty regarding the meaning of various concepts, especially depending on the language used. Therefore, it is advisable to provide a glossary and to require one if none is available. This is important, for instance in cross-professional system design, development, implementation, and management.

The impact of being in command of acronyms should not be underestimated. It involves tacit knowledge that can be crucial for any kind of ICT-oriented work. To illustrate, here are a few acronyms of relevance to the current discussion about data protection:

> AI (artificial intelligence), AGI (artificial general intelligence)[12], ANI (artificial narrow intelligence), BC/AC (before/after coronavirus), CPDP (computers, privacy and data protection), DPbDD (data protection by design and default), DPIA (data protection impact assessment), EDPB (European Data Protection Board), EDPS (European Data Protection Supervisor), FAQ (frequently asked questions), GDPR (General Data Protection Regulation), ICT (information and communication technology), ML (machine learning), NLP (natural language processing), NN (neural networks), PET (privacy-enhancing technologies), PoC (proof of concept), RaC (rules as code), RPA (robotic process automation), SAQ (seldom asked questions).[13]

12 See further, Olle Häggström, Tänkande maskiner. Den artificiella intelligensens genombrott (2021).

13 It is not always that easy to find an adequate acronym. An example of this is the Swedish Authority for Privacy Protection, which in January 2021 changed its name, resulting in the new acronym IMY. In Swedish, this reads just fine – but on social media it might be interpreted as 'I miss you'.

4 The interplay between AI and DPbDD

Here, the focus will be shifted to the combination of and interplay between AI and DPbDD. The major issue remains the same, i.e., whether the conceptual approach seems to add value. This question makes it worthwhile to enter into a short description of the data protection legislation in force and how it is linked to underlying concepts of legal relevance.

An overview of the legal landscape, narrowed down to the EU and a conceptual approach, would highlight the provisions and articles below. The main reason for here inserting quite a lot of GDPR text can somewhat ambitiously be described as pedagogical. More precisely, the purpose is to illuminate the need for a comprehensive approach when it comes legal sources. In practice, it is not enough to consider merely one or two references to governing rules and regulations. Taking the GDPR as an example, someone applying the law needs to consult the following sources: Explanatory memorandum, Recitals, Provisions, Articles and Annexes. In addition to these formally binding components, there are quite a number of other decision-making bodies, e.g., the European Data Protection Board (EDPB), not least taking court cases into consideration.

Yet another reflection is that EU law might come across as quite wordy in comparison to other legal systems within the Member States and internationally. Nor is the logical structure always as stringent as one would expect and wish for. The fact that the normative contents of EU Directives and Regulations are often the results of negotiations is one explanation for this.

It is important to note, concerning the current analysis – as pointed out above – that it is seldom sufficient to focus on any one particular legal definition if it is taken out of context.

> **GDPR**
> Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)

Article 5
Principles relating to processing of personal data
(a) ('lawfulness, fairness and transparency');
(b) ('purpose limitation');
(c) ('data minimisation');
(d) ('accuracy')
(e) ('storage limitation');
(f) ('integrity and confidentiality')

Article 25, (recital 78, 79)
Data protection by design and by default
1. Taking into account the state of the art, the cost of implementation and the nature, scope, context and purposes of processing as well as the risks of varying likelihood and severity for rights and freedoms of natural persons posed by the processing, the controller shall, both at the time of the determination of the means for processing and at the time of the processing itself, implement appropriate technical and organisational measures, such as pseudonymisation, which are designed to implement data-protection principles, such as data minimisation, in an effective manner and to integrate the necessary safeguards into the processing in order to meet the requirements of this Regulation and protect the rights of data subjects.

2. The controller shall implement appropriate technical and organisational measures for ensuring that, by default, only personal data which are necessary for each specific purpose of the processing are processed. That obligation applies to the amount of personal data collected, the extent of their processing, the period of their storage and their accessibility. In particular, such measures shall ensure that by default personal data are not made accessible without the individual's intervention to an indefinite number of natural persons.

3. An approved certification mechanism pursuant to Article 42 may be used as an element to demonstrate compliance with the requirements set out in paragraphs 1 and 2 of this Article.

Recital 78
The protection of the rights and freedoms of natural persons with regard to the processing of personal data require that appropriate technical and organisational measures be taken to ensure that the requirements of this Regulation are met. In order to be able to demonstrate compliance with this Regulation, the controller should adopt internal policies and implement measures which meet in par-

ticular the principles of data protection by design and data protection by default. Such measures could consist, inter alia, of minimising the processing of personal data, pseudonymising personal data as soon as possible, transparency with regard to the functions and processing of personal data, enabling the data subject to monitor the data processing, enabling the controller to create and improve security features. When developing, designing, selecting and using applications, services and products that are based on the processing of personal data or process personal data to fulfil their task, producers of the products, services and applications should be encouraged to take into account the right to data protection when developing and designing such products, services and applications and, with due regard to the state of the art, to make sure that controllers and processors are able to fulfil their data protection obligations. The principles of data protection by design and by default should also be taken into consideration in the context of public tenders.

EDPB Guidelines
European Data Protection Board Guidelines 4/2019 on Article 25, Data Protection by Design and by Default, version 2.0, Adopted on 20 October 2020

Scope concerning Accuracy (78)
The requirements should be seen in relation to the risks and consequences of the concrete use of data. Inaccurate personal data could be a risk to the data subjects' rights and freedoms, for example when leading to a faulty diagnosis or wrongful treatment of a health protocol, or an incorrect image of a person can lead to decisions being made on the wrong basis either manually, using automated decision-making, or through artificial intelligence.

Scope concerning Accuracy (79)
Measurably accurate – Reduce the number of false positives/negatives, for example biases in automated decisions and artificial intelligence.

See further example 1, page 24 EDPB Guidelines:
An insurance company wishes to use artificial intelligence (AI) to profile customers buying insurance as a basis for their decision making when calculating the insurance risk.

From the extracts above follows that the conceptual approach to data protection in an AI setting could be a strong supplement to traditional methods for understanding law (such as legal analytics). However, there are many more modern alternative approaches to be considered. So-called regulatory sandboxes[14] and legal testbeds are merely two examples. These kinds of *legal laboratories*, dealing with hypothetical legal issues in digital proof-of-concept settings and alternative solutions, such as legal standards, are promising and probably necessary. To exemplify, legal requirements on letting data subjects receive meaningful information about the logic behind automated individual decision-making, including profiling (Article 15 1. h) GDPR), must be considered at early stages of cloud computing, edge computing,[15] etc. At the same time, it can be argued that DPbDD (Article 25 GDPR) is easier said than done. Introducing AI is both complex (many components) and complicated (advanced). It is worth mentioning the challenges of conflicting data protection principles, e.g., data minimisation versus big data.

A conceptual approach to AI and data protection commonly requires computer processing power and knowledge of statistics, mathematics, and much more. Governing data models are usually designed to predict outcomes of machine learning approaches. Thus, DPbDD requires means for self-learning (dynamic) algorithms working on training data, validation data and testing data. The combination is a form of autonomous system and software that can hopefully let law be a placeholder.

5 A Nordic law perspective

Of course, this is not the place for a comprehensive introduction to Nordic law, though the interested reader will find a few suitable

14 See for instance the Explanatory memorandum in the EU AI proposal:

'Additional measures are also proposed to support innovation, in particular through AI regulatory sandboxes and other measures to reduce the regulatory burden and to support Small and Medium-Sized Enterprises ("SMEs") and start-ups.'

15 Edge computing can roughly be described as a kind of computing where centralised data processing takes place close to the client, instead of remotely.

references in the footnotes.[16] Instead, three different legal problem areas will be addressed, with a common component in the impact of conceptual models. The first problem area relates to the Swedish principle of openness, while the second concerns public sector records management. The third area – which might strike the reader as a bit odd – is a discussion of what can be referred to as technical interpreters.

5.1 Transparency by way of routine measures

A first substantive example of a normative concept that differs depending on measures taken before or after AI was introduced in a specific public agency relates to the Swedish principle of openness, dating back to the year 1766. Briefly, this right gives anyone, irrespective of legal status, nationality, etc., a right to access official documents that are public, i.e., not secret (confidential).[17]

Digitalisation has had a strong impact on recent developments. While the historical starting point was paper documents, the public sector of Sweden is now almost completely digital. In response to this, law-making bodies have adjusted the scope of the openness principle so that it comprises not only separate electronic recordings, but also *compilations of data*. This is the case when such a compilation can be achieved through so-called routine measures. The notion of routine measures is however conditional in that it presupposes a limited work effort, at reasonable cost, and limitation of any other burdensome actions on the part of the public agency in question.[18] Needless to say, there is a large difference between the routine measures of the past, those of the present day and (probably) those of the future.

16 Ulf Bernitz et al., Finna rätt: Juristens Källmaterial och Arbetsmetoder (15th Edition, 2020) and Rättsinformatik i det digitala informationssamhället (Cecilia Magnusson Sjöberg ed., 2021). Cecilia Magnusson Sjöberg, E-hälsa som app – dataskydd och datadelning (2020).

17 See further Chapter 2, Section 3, Freedom of the Press Act (constitutional law) and the Public Access to Information and Secrecy Act (Swedish Code of Statutes 2009:400).

18 For comparison, under the GDPR:

> '1. Personal data shall be:
> (a) processed lawfully, fairly and in a transparent manner in relation to the data subject ("lawfulness, fairness and transparency")'.

5.2 Records management and big data

Another core concept at the intersection of AI and law is big data, a major feature in today's AI methods. From a privacy point of view, personal data processing in large datasets is challenging.[19] More precisely, it triggers a risk for privacy infringements among data subjects.[20]

The other side of the coin is records management in the public sector of Sweden. Here, the condition for legal compliance is the opposite, i.e., expected long-term storage of (electronic) official documents. The overall purpose is, according to Section 3 of the Archives Act (1990), to satisfy (I) the right of access to official documents, (II) the need for information within adjudication and administration, and (III) the needs of research. It is also considered important to protect the nation's cultural heritage. In summary, legal inclusion of big data in AI systems requires a weighing of interests, between the data protection principle on data minimisation on the one hand and the public's quest for transparency on the other.[21]

5.3 Technical interpreters

A quite innovative approach (comparatively speaking) would be to launch what might be called technical interpreters. The foundation for this concept is existing Swedish legislation on interpretation and translation of natural languages laid down in Section 13 of the Swedish Public Administration Act (2017:900): If necessary, a public authority should, among its other duties, use interpreters and accomplish translation of documents when contacting someone who is not in command of Swedish, thus making it possible for them to claim their rights.

19 For a critical view of digitalisation governed by large tech companies such as Amazon and Google, see Shoshana Zuboff, The Age of Surveillance Capitalism: The Fight For a Human Future At the New Frontier of Power (2019).

20 See Article 5.1 (c) GDPR.

> '1. Personal data shall be:
> […]
> (c) adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed ("data minimisation")'.

21 On e-archives see, e.g., the Swedish Ombudsman JO, Dnr 5698-2014, Decision 20211-05-20. *Kritik mot Kulturnämnden i Stockholms stad för att allmänna handlingar gallrades ur stadens e-arkiv utan att det fanns rättsligt stöd för detta.* (Swedish)

Bridging the gap between national administrative law and AI becomes complicated when applying the provisions on information duties and a data subject's right of access. In particular, it is the wording of Article 15.1 (h) in conjunction with Article 22 that requires attention (see also Recital 63). To get the full picture, national rules and regulations must also be taken into consideration (see above). The concept-related conflict that appears has to do with the GDPR's requirement on data controllers to provide data subjects with meaningful information about the logic involved in automated individual decision-making, including profiling, which could be quite challenging to fulfil.[22] This is where the potential contribution of a technical interpreter enters into the picture. Like in the case of a traditional interpreter, the task would be explanation and translation– but not between natural languages. In a digital environment, it is instead between software (code) and human (natural) language that the interpretation must take place.

6 Concluding remarks

6.1 The DataLEASH project

This brief text has attempted to show the methodological prospects when it comes to a conceptual approach within a digital environment characterised by data processing. In this context, AI plays an increasingly important role as a trigger beyond the traditional legislative means for assessments based on normative rules. Still, rule-based automated measures for legal decision-making will no doubt remain a reality in our societies for many years. At the same time NLP – based on AI – seems likely to become a critical success factor, considering that the law is very much, but not solely, about text.

22 See Article 15.1 (h) GDPR, Article 22 GDPR:

> '1. The data subject shall have the right to obtain from the controller confirmation as to whether or not personal data concerning him or her are being processed, and, where that is the case, access to the personal data and the following information:
>
> […]
>
> (h) the existence of automated decision-making, including profiling, referred to in Article 22(1) and (4) and, at least in those cases, meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject.'

Actually, it seems as if we are experiencing a step towards legal NLP. The reasoning regarding this can easily become somewhat circular with regards to rules and regulations, as concepts and text are regularly changing position in what can be described as a *legal ecosystem*.

As a matter of fact, the future is already here, at least in the form of the *DataLEASH research project* run by a number of Swedish universities in Stockholm, Sweden. The acronym LEASH stands for '**lea**rning and **sh**aring under privacy constraints', and the project opens for letting law play a proactive role instead of the traditional reactive one when matters have already gone wrong. In this context, the GDPR requires attention, in particular Article 5 (principles relating to processing of personal data) and Article 25 (data protection by design and by default).

Among other tasks, DataLEASH will develop and test methods for more open data in a range of disciplines. In practice, the work will consist of risk analyses for *privacy protection* (key indicators, methodology, legal requirements) and *privacy-configured learning systems* (mechanisms, considerations related to integrity and user value). The project will support public organisations that are required to have open data and therefore need data management methods that are fast, reliable and uncomplicated. Based on such methods, they would be able to make well-informed decisions on *how* and *if* data should be shared (limiting access and various security settings) and to choose what form of data conversion that should be aligned with a certain level of privacy and use.[23]

6.2 Forthcoming research

This publication is not the place for an in-depth discussion of what may be referred to as *digital persons*. However, it is interesting to observe that the notion per se attracts attention in many situations related to different (legal) systems. Therefore, a few thoughts will be shared with a focus on forthcoming research. Actually, a few threads of such research have already been presented to the research commu-

23 Read more about the project: *Learning and Sharing Under Privacy Constraints (DataLEASH)*, Digital Futures, KTH Royal Institute of Technology (January 21, 2020), available at https://www.digitalfutures.kth.se/research/collaborative-projects/learning-and-sharing-under-privacy-constraints-dataleash/ (last accessed July 6, 2021).

nity.[24] The major research question is whether there is an innovative legal entity emerging, mirroring a new role of law in an AI-based society.

Historically, society has been populated by *natural persons*.[25] For a variety of reasons, there was a need for something more to meet demands in society and the *legal person* was introduced as an institution of its own. Following from this very broad outline, we are facing a discussion about personhood in the form of a *digital person*.[26]

The question 'why?' or rather 'why not?' appears immediately. To put it simply, if we remain passive when it comes to a new legal personhood, there is a risk for substantive loss. To briefly exemplify: (financial) transaction(s) might not be legally recognised when the intelligent agent making the transaction is not recognised as either a natural or a legal person. This could make business and other activities more complicated and costly than would otherwise be the case.

The reasoning boils down to an open-minded approach to the digital person as a new legal entity. If not accepted and not given certain rights or responsibilities, there is– in a long-term perspective – an increasing risk of a dysfunctional (legal) society. If no one takes the role of being a subject, and to some extent an agent, this raises all kinds of concerns. Common examples of when this might be problematic include self-driving vehicles, pricing algorithms on the competitive market, data protection when profiling consumers, and openness and transparency in digital environments.

If digital persons were to be developed and accepted, there are several critical success factors to be aware of. Not least: technical terms and functions need to be negotiated and standardised, preferably on an international basis. Some core components of AI are predictive modelling based on statistics, mathematics, computer processing power, NLP (including text classification), ML, etc.[27] To summarise, a major distinction ought to be made between conventional automation on the one hand and AI on the other. Digital

24 See, e.g., Cecilia Magnusson Sjöberg, *The Digital Person – A New Legal Entity? On the Role of Law in an AI-based Society*, in Legal Tech and the New Sharing Economy pp. 49–59 (Marcelo Corrrales Compagnucci et al. eds., 2020).

25 In other words: physical, biological human beings.

26 See also Visa A. J. Kurki, A Theory of Legal Personhood (2019).

27 See further Stanley Greenstein, Our Humanity Exposed: Predictive Modelling in a Legal Context (2017).

persons belong to the AI category, as described below, so traditional legal automation will certainly be challenged by AI-based solutions:

> **Automation:** Data collections processed within a system governed by static (deterministic) algorithms and code running on/executed by computers.
>
> **AI:** Processing based on training data, validation data and testing data within models governed by appropriately used, dynamic (self-learning) algorithms.

As a thought experiment, we could imagine the roles of teachers and tutors as digital persons. Such an adaptation would require proactive reasoning and likely be seen as a provocative suggestion. It would also be likely to cause prestige-related reactions within the framework of learning analytics.

The initially presented hypothesis in this contribution has arguably been verified in this study, i.e., a conceptual approach to legally valid data protection creates interesting possibilities. Once again, it should be emphasised that the intention is in no way to replace conventional legal methods, but instead to strengthen the current state of legal analyses. In this context, it is important to focus on functional analyses of applied AI rather than on formal categorisations. Therefore, it is important to bridge the methodological gaps between legacy approaches, current ones and future-oriented ones. In summary: legal tech is here to stay – so it is best to embrace the development towards AI by privacy design and default, acknowledging both the history and the future of law in modern society. Already at this stage, legal infrastructures enabling automated data processing, electronic documentation and communication in global networks need to be in place.

Complexity and Narrative Identity: A Shift from Design to Intention in Privacy Law

SARA GANDRÉN AND NICKLAS BERILD LUNDBLAD

In this article we want to explore a simple question: what happens with the core concepts of privacy as information systems become more complex? Is there a difference in kind or just in degree as to how we should think about data protection in systems that pass a certain level of complexity?

We believe this question is important, on a number of different levels, because of two major reasons: the first is the increased complexity of our technology. As documented by Sam Arbesman in *Overcomplicated: Technology at the Limits of Comprehension* (2017) our technological systems are growing in complexity to a point where they no longer are accessible to the hitherto dominant methods of analysis. The second is that we believe that complexity will lead to something like phase shifts in how we think about privacy.

Data protection law in the future will play out in increasingly complex systems where we need to re-examine old assumptions about privacy, data protection and autonomy.

In the article we first discuss the concept of complexity and why it is reasonable to say that complexity is growing. We then turn to personal data in complex systems, with an emphasis on the question about inferences and if they should be deemed personal data. We go from there to discuss how the objective of data protection may be shifting, and why the so-called autonomy trap is so important. Finally, we suggest a few ways in which it may be necessary to reform data protection law in a world of increasingly complex systems and address some counter-arguments that can reasonably be raised.

On complexity

One of the core assumptions we are making is that information systems are becoming more complex and that this impacts the way we think about data protection and privacy. This is not a self-evident statement, and there are at least two major objections to this argument that we would like to explore up front to lay the foundation for the rest of the argument.

The first is that technological systems' complexity is irrelevant to law, since law deals with behaviours and outcomes, and so the internal complexity of a system is unimportant to the way the system is treated in the external world. After all — law regulates very complex systems, like human beings, already and so there is no reason to think that there would be a material difference in any aspect of the law because of complexity orders of magnitude less than that of the human brain.

The tag line version of this criticism could be "almost no one knows how a fridge works, but we can still regulate them without major concerns".

We believe this is wrong, for a very simple reason: legal analysis has – much like science – depended on the coupling of *comprehension* and *competence*, to use the dichotomy that philosopher DC Dennett employs when talking about artificial intelligence systems and the new conditions they provide.[1]

Our new technical systems are exhibiting human level competence, but they do not have anywhere near the same levels of comprehension. And this means that we also lack the means to comprehend exactly what it is that they are doing. We can follow and evaluate the outcomes, but not the processes whereby we arrived at them. Our knowledge is increasingly devoid of understanding, and the decoupling of understanding and explanation from knowledge is a threshold moment in the history of epistemology.

We are, naturally, aware of the efforts made to make data sets transparent, visualise them and generally achieve explainable artificial intelligence, but our prediction is that this project will fail. While it may be possible to achieve explainability for single isolated

1 See e.g. Daniel C. Dennett, From Bacteria to Bach and Back: The Evolution of Minds (2017).

systems, it seems impossible to achieve that for the sum of interacting systems in a society.

Another way of putting this is saying that we are, in principle, seeing the emergence of a "black box society" where the black boxes need to evolve in two dimensions: first to adapt to the environment they are in, and then, and this is crucial, adapt to each other.[2]

As biologist and philosopher William C Wimsatt has shown, it is this second order evolutionary process in what is sometimes termed a Red Queen's dilemma (as the Red Queen in Alice in Wonderland notes, it takes all the running you can do to remain in the very same place) that creates increased complexity in biological ecosystems, and there is no reason to think that it will not create rising levels of complexity in society as well.[3]

We argue that a network of black boxes is different from a fridge – because of the perpetually increasing complexity such a network will generate, and because of the level of that complexity.

The second objection is that complexity is a *design choice,* and that we should not adapt a human right after a technological development trajectory. We should, instead, prohibit systems that exhibit levels of complexity such that we need to change our concepts of data protection.[4]

Our argument is used in the inverse here: where there is pressure to re-conceptualise privacy or data protection, because of concerns around complexity and explainability, it is much better to stop the evolution of such complexity or prohibit it than to allow data protection to change.[5]

2 The concept of a "Black Box Society" was introduced by Frank Pasquale in his book The Black Box Society: The Secret Algorithms That Control Money and Information (2016).

3 See William C. Wimsatt, Re-Engineering Philosophy for Limited Beings: Piecewise Approximations to Reality (2007).

4 Or, as some argue, adapt our models in various ways to make them more easily interpretable to humans. For a discussion on possible methods to achieve this, see Andrew D. Selbst and Solon Barocas, *The Intuitive Appeal of Explainable Machines*, 87 Fordham L. Rev. 1085, 1110–1115 (2018).

5 See for a longer discussion about this, with a grounding in i.a. UK discussions, Michèle Finck, *Automated Decision-Making and Administrative Law*, in Oxford Handbook of Comparative Administrative Law (P. Cane et al. eds., 2020).

This is, logically, a possible legislative position, and there are some who have argued that this is a tenable solution. We are less sure of that. One reason we believe this is wishful thinking is that our systems may already have entered the phase where this is happening – which is why questions of inferences as personal data, to take one example, are of increasing importance. We already live in a society where complexity is creating black boxes, and prohibiting black boxes would mean rolling back not one or two AI-systems, but many more systems.

Another reason is that complexity is not a net negative. It seems possible that these systems will not just match, but surpass, human competence in many areas, like medicine. That means that if we prohibit complex systems or black boxes, we risk denying ourselves the ability to solve key problems facing us. *Complexity is tied to capability.*

Overall, our view is that complex systems will not be prohibited, but that legislation and the political debate will first focus on transparency and explainability (as it already is), but then need to find new solutions and new approaches as we run out of runway for those two measures.

That is why we believe we need to revisit basic concepts of data protection and privacy and examine what happens with them in a complex systems environment – if nothing else to explore possible reform paths. We begin with an examination of a current hot topic: the right to reasonable inferences.

Should there be a right to reasonable inferences?

For as long as the General Data Protection Regulation[6] has existed, scholars have argued over whether its article on automated decision-making constitutes a right to explanation.[7] Sandra Wachter

6 Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (hereinafter "GDPR").

7 See, e.g., Margot E. Kaminski, *The Right to Explanation, Explained*, 34 Berkeley Tech. L. J. 189 (2019); Lilian Edwards & Michael Veale, *Slave to the Algorithm? Why a 'Right to an Explanation' Is Probably Not the Remedy You Are Looking For*, 16 Duke L. & Tech. Rev. 18-84 (2017); Andrew D. Selbst & Julia Powles, *Meaningful Information and the Right to Explanation*, 7 Int'l Data Privacy L. 233-242 (2017); for an excellent

has, together with other authors like Luciano Floridi and Brent Mittelstadt produced a large body of work on inferences and data protection law, and argues in a 2019 paper for a right to reasonable inferences.[8] The idea is clear: when systems make inferences about us those inferences should a) be classified as personal data and b) be subject to some conditions of "reasonableness".

Wachter is right in arguing that inferences are becoming more important, and that the current data protection law is not equipped to deal with them. As she has shown, the European Court of Justice has consistently taken a bleak view of the idea that inferences are personal data, and if they are they have employed a teleological interpretation of data protection law to exclude inferences from the system of rights and duties in the law. Wachter's argument is that this weakens data protection law in an unacceptable way and leads to a situation where we have no right to reasonable decisions being made about us, since we cannot govern what inferences are permissible and accurate in making those decisions.

If we examine Wachter's proposal for a right to reasonable inferences from a complexity perspective we find several interesting challenges.

First, the reasonableness criterion becomes hard to explicate: what does it mean for an inference to be "reasonable" if the inference is produced by an opaque process? Wachter's reasonableness test looks at three discrete steps: why certain data form a normatively acceptable basis from which to draw inferences, why these inferences are relevant and normatively acceptable for the chosen processing purpose or type of automated decision and then whether the data and methods used to draw the inferences are accurate and statistically reliable.

In addition to this ex ante justification mechanism Wachter suggests a right to contest inferences that have been made about a person ex post.

---

discussion on the various examples and perspectives of these arguments, see also Selbst & Barocas, *supra* note 4.

8 See, for example, Sandra Wachter, Brent Mittelstadt & Luciano Floridi, *Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation*, 7 Int'l Data Privacy L. 76-99 (2017); and specifically Sandra Wachter & Brent Mittelstadt, *A Right to Reasonable Inferences: Re-Thinking Data Protection Law in the Age of Big Data and AI*, 2 Columbia Bus. L. Rev. 494 (2019).

There is a lot to recommend Wachter's model, especially for those that are comfortable with privacy and data protection law slowly transforming into decision making standards. The argument put forward here is essentially an argument for what Wachter et al call a right to be seen in a fair way – and we could develop this further, and note that the trajectory that Wachter sketches here is one in which privacy is protected by allowing greater control over the way identity is produced.

It is a right to one's own narrative, in a sense.

But the idea of a right to reasonable inferences seems to be premised on *explainability*. If we cannot explain what data has been used or how the inferences have been made, then the idea of a right to reasonable inferences falls apart not as undesirable, but as impossible.

The assumption of explainability may seem a modest requirement, but when we look at the way inferences are drawn in more and more complex systems it is not clear that it is possible to use the justification mechanism that Wachter et al propose. Now, a simple way for Wachter to reply to such criticism would be to say that it is normatively unacceptable to use data or methods that are not explainable. The normative criterion in her justification mechanism opens a vast array of possible extra restrictions on inferences, but it seems far from clear that explainability is, or at least will remain, a key norm in our societies.

We should also address another weakness in the argument for reasonable inferences, and that is that it seems as if the word "inference" is made to do heavy lifting here. Inferences are key elements in decision making and reasoning, but it is never quite clear if they have to be made by human beings or if a machine or a piece of software can make an inference.

It is not too far-fetched to argue that inferences are intentional in the sense that only people can draw them, and what they then use is clear correlations or causal patterns suggested by data analytics (can we really talk about causal patterns? There is a growing body of work that suggest that we can, and should, reintroduce causality into the conversation – se for example the work of Judea Pearl).

The right to reasonable correlations seems to be a better way to limit the justification to the system used, than the idea that there is a right to reasonable inferences on the basis of the identified correlative pattern.

One reason this distinction matters to us is that if we agree that inferences are drawn by human beings, then we have already introduced one black box in the system, and it should not be impermissible to introduce a second or third, even if it is in software.

Humans are, by all measures, complex systems. It is not always easy to explain human reasoning – even if we can motivate and defend our reasoning. Often we use psychology in order to understand human behavior and actions, and psychology is a science that deals almost entirely with actions and decisions, and from them try to construct explanations – always acknowledging that these are provisional explanations.

When we argue that a right to reasonable inferences will be hard to introduce, because the process is opaque and increasingly complex, we are essentially arguing that it is impossible to legislate rationality – from either machines or humans. There will always be significant amounts of noise and arbitrary variation in human decisions, and regulating the systems providing correlation patterns and data is not going to help that.

This also complicates the question of whether inferences are personal data. We would suggest that inferences are ordered in narratives about individuals, and so represent a kind of fictional personal data, whereas simpler data can be described as factual personal data.

The question of if inferences is personal data then becomes a question about whether fictional personal data is really personal data. If I tell a story about you is that story then a set of personal data? If we do indeed think that the free production of identity, coupled with narrative control, is the essence of privacy protection, then we should argue that fictional personal data is personal data and should be subject to all the rules in data protection law.

Doing so highlights the age-old conflict between privacy and free speech – since the essence of free speech is the right to tell stories about the world and the people in it.

From the perspective of complexity, however, we would note that there is no difference between a human being and a machine telling stories about us. Both are opaque systems, black boxes, that ultimately work in ways that we cannot explain in any detail.

Inferences are narratives, and narratives are already today produced by black boxes, and it is hard to imagine a right to reasonable stories, which is what Wachter et al ends up arguing for. We see

this more clearly if we realize that the justification mechanism that Wachter suggests fails under conditions of opaqueness and complexity.

Interestingly, the Federal Administrative Court of Austria recently decided[9] that a natural person's affinity for a political party, inferred from, e.g., regional election results, opinion polls and socio-demographic information, constitutes personal data and that such data is capable of being rectified. The complainant in the case had argued, inter alia, that probability values lack any statement or information regarding a natural person, and that the right to rectification does not apply to probability values. If uncorrectable data were to be deemed as personal data, the complainant argued, it would create a subset of personal data type to which all data subject rights would not apply, which would be a concept foreign to the GDPR.

The Court however referred to the three-step test proposed by the Article 29 Working Party as a way to determine whether data is personal data,[10] stating that the inferences regarding political affinity in the case at hand fulfilled all three steps. In particular, this was true in regards to the results of the inferences, which ultimately lead to different natural persons being treated differently depending on their inferred political affinity.

Regarding the objection that probability values cannot be rectified and thus cannot be defined as personal data, the Court rejected it out-right, stating that even if such data could not be rectified, that would not mean that the GDPR would be inapplicable in its entirety. Furthermore, the Court argued that the accuracy of personal data must be assessed in relation to its purpose for processing; since the complainant's purpose was not to determine the data subjects' political affinity in concrete terms, but only to make a statistically sound assessment of such affinity, the data was deemed to be correctable in that errors of the assessment – such as the use of incorrect base data – could be rectified. Thus, since a new assessment based on more correct data could give a new and correct determination of party affinity, the Court established that the inferred political affinity could indeed be rectified.

9 Bundesverwaltungsgericht (BVwG), case number W258 2217446-1, 26.11.2020.

10 Article 29 Data Protection Working Party, *Opinion 4/2007 on the Concept of Personal Data*, European Commission, 01248/07/EN WP136, (June 20, 2007).

The Court's conclusions lean somewhat in the direction of Wachter's proposed model regarding the right to reasonable inferences, in that data controllers would need to ensure that – and explain why – the data behind the inferences are accurate and statistically reliable. However, while the right to rectification in this case was extended to the data behind the assessment, the Court did not mention any right to rectify the assessment itself. And again, the opaqueness of more complex systems begs the question whether this is even possible. If a machine makes inferences that we humans cannot fathom, how can we even begin to identify which part of the assessment needs rectification in the first place?

In their paper on the role of explanation for AI accountability,[11] Finale Doshi-Velez et al suggest another plausible solution to the dilemma of complexity and explainability by arguing that *explanation* does not necessarily equal *transparency.* They argue that if any information is to be useful, the correct type of information must be provided. As a result, they propose that an explanation should be able to provide human-interpretable information about at least one of the following: which factors went into the decision, or which factors were determinative of a specific outcome.

Considering the growing complexity of machine learning, we argue that both of these properties can be satisfied without ever knowing the fine details of how the AI reached its decision: We do not need to know the flow of bits through an AI system any more than we need to know the flow of signals through neurons (both of which would be uninterpretable to a human, at any rate).

Autonomy and complexity

Wachter's valuable work also highlights another trend: the increasing focus on decision making in data protection law. Standards for decision making are increasingly important and just as we have accepted that we need, for example, sentencing guidelines, it seems

11 Finale Doshi-Velez et al., *Accountability of AI Under the Law: The Role of Explanation*, Berkman Center Research Publication (Forthcoming) (November 3, 2017), available at SSRN: https://ssrn.com/abstract=3064761 or http://dx.doi.org/10.2139/ssrn.3064761.

reasonable to accept that we will need decision making standards of different kinds.

Most of the debate about decision making standards have to do with decisions being made about us by others, but it can be quickly recognized that if the objective, or one objective, of data protection is to protect our autonomy and right to self-determination, then that same concern should be extended to the standards we use in making decisions about ourselves.

This first comes off as slightly counter-intuitive, since the right to autonomy seems to be the right to pick whatever standards we want when we make decisions about ourselves, but things are not quite that simple.

Very few people are completely at liberty to pick whatever standards they want when they evaluate themselves. We use normative structures, the praise or castigation we experience from others, peer pressure and generally simplified models to understand ourselves. Not to mention that we are notoriously bad at making both decisions and predictions, yet we constantly overestimate our abilities in those very areas. We are biased in a plethora of ways, most of which are difficult to shake even after we become aware of them. And ultimately, we simply don't possess the mental resources to deliberate on every single decision in our lives. The great attraction of personality tests like Myers Briggs et cetera is not just because it makes it possible for us to understand others, but because it makes it easier to understand ourselves – in the frames that are offered by the tests (whether accurate or not!).

Why, then, is this relevant to the impact of complexity on privacy? The reason is simple: imagine that you knew that a very complex system, with a good track record, made inferences about you that said that you were likely to change careers and become a priest – how would you then use that signal in your understanding of yourself?

The system has made a prediction about you, and knowing that this is a complex system with superhuman abilities, you are likely to integrate the system's view of you in your own decision making standards about yourself. You now face a peculiar choice – either you conform to the prediction, or you decide that you will consciously deviate from it. In either case you have lost your autonomy – you are now reacting to a decision making standard informed by a machine and a complex system that you cannot explain or understand.

In fact, the complexity of the system could mean that we ourselves may not even be able to determine the correctness of an inference drawn about us. When we move away from inferences about facts that we already know about ourselves, and towards predictions about how we might behave in the future, it becomes easier to simply trust the machine than to evaluate whether the conclusion it has drawn seems correct or not.

This results in a kind of *autonomy trap*, where your decisions are now decisions made under the gravity of the prediction offered by the machine. It gets worse as the machine gets better – if it is able to predict what is good for you and what will make you happy with greater and greater accuracy you are left with the free will to make slightly worse choices.

As our systems grow more complex we should also expect them to improve in quality – the only reason to increase complexity in a system deliberately is if we believe that it can perform better. That means that with increased complexity we should also expect to see increased accuracy in predictions and descriptions of individuals.

Paradoxically this could mean that if we increase transparency, in the sense that individuals get more access to what these systems infer about them, then we reduce individual autonomy as individuals become more and more beholden to the predictions and descriptions produced by the systems as such. And knowing what the systems have as in data and the methods employed will do little to lessen this effect if the data sets and methods are increasingly complex.

Autonomy is undermined, then, by any increased access and transparency to, and with, the complex systems that make inferences about us, the very means that we had hoped would help us avoid such effects from automated decision making. This is not said to deny that there is very real value to transparency in a large set of cases, it is just to point out that increased transparency into a predictive system can have deleterious effects on perceived autonomy.

An interlude on intentional terms and agency

Before we move on to exploring what this will mean for future data protection reform, we need to say a few words about intentional terms and how they relate to and confuse this debate.

Wachter et al speak of systems and machines as making inferences, and there is a tendency in the literature to speak of automated decision systems.[12] Most of this literature passes over in silence the question of agency and intention, even though it is a fundamental question for us to understand if we want to discuss decision making standards or inferences.

The systems we are discussing here are not legal subjects in the sense that they have the capacity to act in legal ways. The use of these systems presupposes a human being with agency that at a very minimum decides that these systems shall be used in order to inform the decision that they are making.

Only human beings make inferences, draw conclusions or make decisions. Machines do not.

If we look closer at all the questions around automated decision making or inferences, we find that there is always a human in the loop at some point – at the point of installing the systems or deciding that they are adequate for the purpose they are deployed. If we ignore this we are allowing for a strange phenomenon that we can call "agency creep" where we push the responsibility for the decisions and inferences from the individual deploying or installing the system to the system itself.

We turn a question of individual accountability for the use of technology into a technology design or access question. If we say that systems must be explainable for us to be able to use them, we are suddenly requiring that the system designer solve the problem that the system user should really be held accountable for.

This is easy to realize if we introduce a quick thought experiment with a world in which these machines have become ubiquitous and anyone can consult a learning system of some sort for any interpersonal decision – we could even imagine that such systems have been implanted in our brains and that no one knows if we consult them or not when we make decisions.

12 See e.g. Bryan Casey, Ashkon Farhangi & Roland Vogl, *Rethinking Explainable Machines: The GDPR's 'Right to Explanation' Debate and the Rise of Algorithmic Audits in Enterprise*, 34 Berkeley Tech. L. J. 143 (2019). For a discussion on AI as "prediction technologies" and the impact of such technologies on decision-making, see Ajay K. Agrawal, Joshua S. Gans & Avi Goldfarb, *Prediction, Judgment and Complexity: A Theory of Decision Making and Artificial Intelligence*, NBER Working Paper No. w24243 (2018), available at SSRN: https://ssrn.com/abstract=3112030.

We can choose to ask the system if it thinks we should trust someone or not, and the system will suggest an answer to that question that we then can do what we will with. The whole process is hidden from the other person, and it would make no sense to then argue that these systems should be designed in a special way.

If we do that then we put the systems themselves out of sight and reach, and what then remains is the individual accountability for the decisions we make – without any attempts to regulate technology that we cannot access or force into a certain design anyway.

What we instead end up with is a world in which we focus entirely on decision quality and not on personal data at all. It does not matter what data was used in producing a correlation that I may or may not act on. What matters if I act on it in a way, and in a pattern, that can be said to discriminatory or flawed in some other way.

Following the train of thought that starts in the discussion about automated decision making and inferences, with added complexity and opacity, leads us to a shift from the use of data to the making of decisions, and that is a shift that requires careful democratic discussion before we undertake to support it.

Future questions for data protection reform

We have suggested in this short article the following: technology and information systems are increasingly becoming so complex that our efforts to make them explainable will fail. At the same time they become more and more accurate and capable of solving harder and harder problems, and so the value they bring will increase. This in turn means that data protection legislation is an example of *misleveled legislation,* legislation that is targeting one level in a system – in this case the algorithmic one where the use of data is the relevant unit of analysis – rather than the one where the legal effect can be discerned.

One way to think about this is to use a framework from the aforementioned philosopher DC Dennett. In his seminal work *The Intentional Stance* (1987), Dennett suggests that we are applying different stances as we explain and understand different kinds of systems. Dennett's levels are the physical stance, where physics allow us to explain a phenomenon, the design stance, where we assume the design of a system and understand it almost mechanistically and

the intentional stance, where we assume intentions and use psychological terms and models to explain how a system acts and works.[13]

Data protection and privacy law has been focused on the design level, rather than the intentional level, and so is increasingly collapsing under the complexity of the explanations and models needed to regulate what is increasingly becoming much more economical and efficient to explain as an intentional system.

Note that this does not mean that we are saying that the system is intentional on its own – but the compound of the individual using the system and the system should be seen as one single intentional system and so be regulated as a single system as well.

That is why the decoupling of the system and the user, and the ascription of mental and intentional terms to the technological system or designed system, is a flawed model; we end up confusing the kinds of legislation we need and the kinds of solutions that we should look for.

Indeed, one outrageous solution would be that we need to give up all of data protection legislation as design legislation and instead start exploring what intentional legislation will look like as these systems become more and more intensely fused with human agency.

This outrageous conclusion would naturally be resisted by anyone who still believes that the right level of regulation of these systems is the design level, and this is understandable. But it should be clear that as the design becomes a) increasingly complex and b) closer tied to human agency, the resulting efficacy of design legislation will lessen to the point where it will be but unrealistic and harmful in that it does not catch the real values at stake, nor protect the rights we care about.

Indeed, one way of reading Wachter et al is to say that they are proving beyond doubt that the European Court of Justice, when it refuses to apply data protection law to guarantee decision quality or accuracy, has run up against the boundary of design legislation and that this in itself is an argument not for amending that legislation through the introduction of new rights on the design level, but for looking at intentional models and accountability solutions.

Accountability, liability and responsibility for actions overall and decisions in particular seem a better place to start than to continue

13 Allan Newell, one of the early AI-pioneers, similarly spoke about device level, program level and knowledge level.

a more and more beleaguered discussion about design legislation focused on algorithms, data sets and code. Or to speak with Lessig and his model of the four regulators: over time code fades into markets and norms.[14]

This becomes clear when we see some of the more inventive and exciting solutions proposed for dealing with the impact of machine learning and AI on data protection and decision quality. In Mireille Hildebrandts brilliant article about the rise of agonistic machine learning, we see that what she is describing is nothing else than the adversarial process, recast in design terms. But why translate a perfectly functioning intentional process with a defendant or ombudsman into code, when we do not need to? If our belief is that decision quality is improved by using an adversarial process, we should recommend that and not recommend the design of an inferior system that accomplishes a part of what a human, or a human with a system could do.[15]

Likewise, while it may be technically feasible to create AI systems that are capable of providing the same level of explanation as humans are expected to, capable of and required to provide under the law, Doshi-Velez et al urge us not to mindlessly go down that path. Creating AI systems in our image would, in this case, require an unnecessarily large number of resources that might end up disadvantaging less-resourced companies, resulting in suboptimal – although easily explainable – models.[16]

So, in concluding then, we propose the following: privacy will remain a human right of great importance, but design legislation solutions like the current data protection law will fade away and be replaced by a general right to one's own narrative protected through the allocation of intentional accountability and responsibility much like the one we see in publishing and press legislative frameworks, but with a focus not on publishing as much as on a much broader category of decisions.

14 Lawrence Lessig's famous model of the four regulators is found in Lawrence Lessig, Code and Other Laws of Cyberspace (1999).

15 See Mireille Hildebrandt, *Privacy As Protection of the Incomputable Self: From Agnostic to Agonistic Machine Learning*, 20 Theoretical Inquiries in Law 83-121 (2019).

16 See Doshi-Velez et al., *supra* note 11.

The way to protect privacy, and autonomy, is not through a more complex regulation mirroring the complexity of the design of the systems we rely on, but through a shift of stances.

Part 2

Transparency

Is There a Human Right to Human Contact? Preliminary Reflections on the Robotization of Caregiving

KATARINA FAST LAPPALAINEN

1 Introduction

A single man in his 60s met his fate in his apartment in Oslo in the spring of 2011. He had been married several times and had children.[1] In spite of this, he was found only after more than nine years, by a janitor. Meanwhile, his bills had been paid automatically. In 2018, his pension was stopped since the Norwegian Labour and Welfare Administration could not get in touch with him. According to Arne Krokan, professor at the Norwegian University of Science and Technology, this tragedy would have been unlikely 30 years ago. It is 'the price we've paid for digital services'. Our technological systems simply do not raise any red flags when someone does not make physical contact.[2]

This tragedy raises important questions about the digitalisation of society and the loneliness and social isolation that can result, especially among older individuals, how technology may contribute to such situations, but also how technology could be used to avoid them.

1 Special thanks to Stanley Greenstein, Liane Colonna and Paul Lappalainen for valuable comments.

2 *Man's body was found after lying in Norway flat for nine years, says police – Oslo death sparks questions about role of technology in reducing physical contact in society*, The Guardian, 9 April 2021, Helen Livingstone.

Modern welfare states, such as the Scandinavian states, are facing considerable challenges due to changing demographics, aging populations and decreasing birth rates. In 2020, the proportion of people over 65 years was around 20 percent in the EU, with people over 80 years making up nearly 6 percent of the population, a number which is expected to increase.[3] An aging population and greater longevity means higher rates of people with chronic and/or multiple diseases and various age-related disabilities in need of long-term care. At the same time, there are not enough care workers to provide the care needed and informal care from family or relatives is often not available. Furthermore, there are increasing mental health issues within this part of the population due to involuntary social isolation and chronic loneliness. On top of these dark prospects come the economic aspects of this development, which are worrisome to say the least.[4]

The introduction of personal care robots in caregiving, to provide not only physical but also emotional support, is gaining traction as an attractive and cost-effective part of the solution to these problems. Some even go so far as to claim that there is a need for a 'gigantic technological shift' to tackle the demographical challenges we face.[5]

Nevertheless, the use of personal care robots in caregiving raises many different ethical and legal dilemmas regarding privacy and personal data, as well as concerns in relation to human dignity and the prohibition of inhuman and degrading treatment. At a more

3 Eurostat: Statistics Explained, *Population structure and change*, https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Population_statistics_at_regional_level#Population_change.

4 These issues have recently been raised by the EU Commission in a Green Paper of 27 January 2021, *Green paper on ageing – Fostering solidarity and responsibility between generations*, COM (2021) 50 Final, https://ec.europa.eu/info/sites/info/files/1_en_act_part1_v8_0.pdf. Murthy, the former Surgeon General of the United States, has described the 'loneliness question' as an urgent public health issue, not least because of the digitalization of society, see Vivek H. Murthy, Together – Loneliness, Health and What Happens When We Find Connection, Wellcome Collection (e-book), 2020; Jacob Sweet, *The loneliness pandemic: The psychology and social costs of isolation in everyday life*; Harvard Magazine (2021), https://harvardmagazine.com/2021/01/feature-the-loneliness-pandemic.

5 Rose-Marie Johansson-Pajala et al., *Care Robot Orientation: What, Who and How? Potential Users Perceptions*, International Journal of Social Robotics (2020) vol. 12, p. 1104.

fundamental level, this raises two central issues. First, the question is whether an individual can claim a right to human contact and to be cared for by humans, in an increasingly digitalized and robotized society. If so, a second question arises: on which legal grounds, and to what extent, can such a right be asserted in relation to robotized caregiving? What would the ethical and legal consequences be if the care of humans were turned over to robots for long periods of time? The purpose of this paper is to provide certain reflections on the legal grounds for a possible right to human caregiving and related issues, which as a general legal framework[6] could affect the way in which care robots can be introduced into caregiving. In other words, if there is a right to be left alone, is there also a right not to be left alone?

2 Care robots

Care robots have been under development for some time and can already carry out many of the tasks of care workers. Personal care robots can be defined as non-medical robotics that are created to improve the quality of life of humans.[7] Physically Assistive Robots (PARs) are constructed to perform physical tasks and can carry out daily chores such as getting food and drinks, taking care of personal hygiene, heavy lifting, and transportation. There are also Socially Assistive Robots (SARs), more commonly known as 'social robots', which have been developed to take care of the social and psychological needs of individuals. They can perform therapeutic and cognitive tasks, and provide companionship and simple social entertainment, such as playing games.[8]

6 On the role of fundamental rights in AI, see Proposal for a Regulation on a European Approach for Artificial Intelligence, laying down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending certain Union Legislative Acts, Com (2021) 206 Final, p. 10; High Level Expert Group on AI of the EU (AI-HLEG) in *Ethics guidelines for Trustworthy AI,* European Commission 8 April 2019, p. 9 f.

7 *Ibid.*, p. 1103.

8 Lillian Hung et al., *The benefits of and barriers to using a social robot PARO in care settings: a scoping review*, BMC Geriatrics (2019) vol. 19, p. 232 f.; Rosalie Wang et al., *Robots to assist daily activities: views of older adults with Alzheimer's disease and their caregivers*, International Psychogeriatrics (2017), vol. 29(1), pp. 67–79.

At present, personal care robots are still costly and mainly used by care institutions that are publicly funded. They are not yet commercialized for a mass market at affordable prices. A recent study showed that only 6 out of 107 developed care robots were commercialized.[9]

However, it has been argued that there is a pressing need for commercialization, since public funding appears insufficient to cover the growing demand for care services. Furthermore, the generations currently aged between 50–70 years, the so-called 'baby boomers', are to a considerable extent expected to be able pay for their own devices as well as be more amenable to self-management.[10] Nevertheless, this poses a great risk of inequality between those who can afford to buy a care robot and those who cannot. Should this be a common right?[11]

Other challenges are related to public knowledge of care robots, which is scarce and often provokes negative associations.[12] There is of course the issue of problems of not having a 'human-in-the-loop', i.e., human intervention in every decision cycle of a system,[13] which is illustrated in the case with the lonely Norwegian described above. Possible solutions to this problem, albeit not necessarily through physical contact, can be found in cloud computing, which can 'enable the systematic use of multiple heterogenous devices at different facilities'.[14] This makes the case for not only Healthcare as a Service but also Care as a Service.

One of the most noted examples of social robots for therapeutic purposes is PARO, a baby harp seal robot, created as pet therapy for persons with dementia. It is a well-known fact that real pets are beneficial in the support of persons with this disease. On the other hand, real pets need care, can bite, and may cause allergies. A pet robot is therefore a more practical alternative in this context. There is a lim-

9 Johansson-Pajala et al., *supra* note 5, p. 1104.

10 Gillian Ward et al., *Developing assistive technology consumer market for people aged 50–70*, Ageing & Society (2017) vol. 37, pp. 1050–1067; Tim Blackman, *Care robots for the supermarket shelf: a product gap in assistive technologies*, Ageing & Society (2013) vol. 33, p. 763.

11 Johansson-Pajala et al., *supra* note 5, p. 1110.

12 *Ibid.*, p. 1105.

13 High Level Expert Group on AI of the EU (AI-HLEG) in *Ethics guidelines for Trustworthy AI*, European Commission 8 April 2019, p. 16.

14 Ricardo De Mello et al., *On Human-in-the-Loop CPS in Healthcare: A Cloud-Enabled Mobility Assistance Services*, Robotica (2019) vol. 37, p. 1478.

ited number of research studies available on the effects of PARO. Nevertheless, there is enough evidence to indicate that PARO in most cases reduces negative emotions and behavioural symptoms, while promoting social engagement and as a consequence reducing use of psychotropic medication.[15] The negative effects of PARO were mainly related to the risk of infection due to PARO's fur, which is difficult to clean, as well as ethical issues, such as the risk of infantilising and dehumanizing care.[16]

It is noteworthy that the discussion about care robots mainly concerns the care of the elderly and disabled, but can also extend to other state-run institutions such as schools, prisons, and asylum detention centres.

3 Legal aspects of care robots and human contact

There is certainly a need for a 'technological shift' to cope with the demographical challenges ahead of us. As stated previously, it is relevant to discuss whether it ought to be a right for all citizens of the welfare state to receive assistance from a care robot,[17] which could give rise to a legal right to such care, possibly on the basis of the principle of human dignity.[18]

In this paper, another side of the coin will be discussed, which is related to the risk that persons in need of care are left to the care of robots due to budgetary reasons and a lack of care workers, becoming more or less isolated with little or no human contact. This process could be sped up, for example in the case of a future pandemic. This more dystopian prospect is likely to put a damper on the enthusiasm for care robots, but it is necessary to take this into

15 Hung et al. (2019), *supra* note 8, p. 235 f.

16 Lexo Zardiashvili and Eduard Fosch-Villaronga, *"Oh, Dignity too?" Said the Robot: Human Dignity as the Basis for the Governance of Robotics*, Minds and Machines (2020), p. 130 f; Hung et al (2019), p. 236.

17 Johansson-Pajala et al, *supra* note 5, p. 1110.

18 Catherine Dupré, *Art. 1* in The EU Charter of Fundamental Rights – A Commentary (Steve Peers et al., eds.), Hart Publishing 2014, p. 17 (01.31); Brownlee has suggested a right against social deprivation consisting of 'minimal adequate opportunities for decent and supportive social contact', which comprises both negative and positive aspects, see Kimberley Brownlee, *A right against social deprivation*, The Philosophical Quarterly, Vol 63, No 251, April 2013, p. 207.

consideration when making regulatory choices for the future.[19] How can and should we use care robots and to what extent? What are the legal limits? Is there a right to human contact?

Currently, there is no known legislation that explicitly and specifically regulates either the use of care robots or the right to human contact. Care of the elderly, the disabled and children is for the most part carried out by human care workers. The quality of this 'human care' might vary due to lack of time, funding, shortages of staff or abuse, but human contact as such is not really the issue. This is presumably about to change.[20]

3.1 The study of rules and tools in interaction

In view of the robotization of caregiving, the lawyer has to undertake an analysis of law by examining constitutional and human rights as well as the existence of legal principles related to human contact within the legal system. This calls for an intradisciplinary point of view.[21] Comparisons can be made to other fields of law where issues regarding human contact are generally dealt with, such as penitentiary law, mental healthcare law and migration law, where issues regarding human contact are present in decisions regarding solitary confinement, visits from loved ones, and geographical placement. Parallels can also be made to animal law, where the right for animals to exhibit natural behaviour includes the right to live alone or in a group, depending on the species.[22]

Moreover, the current coronavirus pandemic, with long-lasting lockdowns in many countries and bans on gatherings and visits to nursing homes, will surely give us examples of what can and cannot

19 Stanley Greenstein, *Elevating Legal Informatics in the Digital Age* in Sonya Pettersson (ed.) Digital Human Sciences: New Objects – New Approaches, Stockholm University Press (2021) p. 161 f; Zardiashvili and Fosch-Villaronga *ibid.*, p. 137 f.

20 Allwood for example states that 'The drift seems to be that all human services, that can be digitalized and replaced by computer program-based services, are disappearing.' Jens Allwood, *Is Digitalization Dehumanization? Dystopic Traits of Digitalization*, Proceedings of the IS4SI 2017 Summit – Digitalisation for a Sustainable Society, p. 2, https://www.mdpi.com/2504-3900/1/3/259.

21 Ahti Saarenpää, *Legal Informatics: A Modern Social Science and a Crucial One*, Scandinavian Studies in Law (2018) vol. 65, p. 23.

22 Birgitta Wahlberg, *Animal Law in General and Animal Rights in Particular* in Animal Law and Animal Rights, Scandinavian Studies in Law (2021) vol. 67, p. 31.

be considered necessary and proportionate in a democratic society once these measures have been evaluated and the remedies through legal challenges in court have been exhausted.[23]

In this paper, the focus is primarily on the human rights aspects of social robots and the right to human contact. This relates to human dignity, protected under Article 1 of the EU Charter on Fundamental Rights (EUCFR), which came into force in 2009, and more specifically to the prohibition of torture and inhuman or degrading treatment under Art. 4 of EUCFR and Art. 3 of the European Convention of Human Rights and Fundamental Freedoms (ECHR) of 1950. The case law of the European Court of Human Rights (ECtHR) is of particular importance in this regard, as the case law concerning the EUCFR is very limited in this field. International human rights law will also be accounted for to the extent that it is relevant to this analysis.

Human rights concepts such as dignity and inhuman and degrading treatment are however broad, vague and designed to be used for all kinds of situations as an 'umbrella' or framework and require interpretation in a myriad of different situations.[24]

To this end, it is necessary to begin with elaborating on how human contact can be defined. There seems to be no legal definition established, which means that other fields of research have to be consulted, such as medicine and psychology, particularly the scientific branch known as the science of touch and anthropomorphic phenomena.[25] In other words, an interdisciplinary perspective is necessary in this regard.[26]

23 There have been several challenges in the courts mainly due to freedom of movement and the right to conduct a business. See, for example, a Dutch case, where a lockdown was found to be unconstitutional: *Covid: Dutch crisis as court orders end to curfew,* BBC News (February 16, 2021), https://www.bbc.com/news/world-europe-56084466.

24 Aharon Barak, Human Dignity – The Constitutional Value and the Constitutional Right, Cambridge University Press 2015, p. 157; Robert Alexy, A Theory of Constitutional Rights, Oxford University Press 2002, p. 233.

25 See for example, Tiffany Field, Touch, 2nd edition, MIT Press 2014.

26 Greenstein, *ibid.,* (2021) p. 167 ff, p. 174, p 177; Susan M. Sterett, *What is Law and Society?: Definitional Disputes* in The Handbook of Law and Society (Austin Sarat & Patricia Ewick eds.), Wiley Blackwell 2015, p. 12–13; Peter Seipel, *IT Law in the Framework of Legal Informatics*, Scandinavian Studies of Law (2004) vol. 47, p. 32 f.

The emphasis in this paper is on the interaction of rules and tools and their interdependency, which encompasses consequences both for the individual and for society as a whole.[27]

Moreover, the perspective is the future. In order to assess which regulatory choices regarding the use of care robots might be appropriate from both a legal and ethical standpoint, a forward-looking and proactive perspective is needed.[28] The method is in this respect a tool for 'technological risk management'.[29]

3.2 Human contact

There are many ways that humans can communicate in society that can be meaningful to their existence, even at a distance through traditional letters, post cards, phone calls, e-mails and, perhaps more importantly, video calls. However, human contact is certainly not only about social interactions via various media.

At the heart of human contact is physical contact. Physical contact is deemed a basic human need, like thirst and hunger. The scientific evidence for the importance of human touch shows that it is essential to human health. The Romanian orphanage crisis revealed in the 1980s is a horrific example, with numerous children found tied to their beds where they spent most of their time with little positive physical contact. The neglect and abuse of these children had severe medical and mental health consequences for them.[30]

Touch deprivation increases stress, which in turn increases the levels of cortisol in the brain, potentially causing the affected individual to enter survival mode. Moreover, touch deprivation can lead to an increased heart rate, high blood pressure, increased respiration, muscle tension and sleep deprivation, which in turn can supress the digestive and immune systems, putting the individual at greater risk for infection. It is often linked to cardiovascular diseases and even

27 Greenstein, *ibid.,* (2021) p. 156; Seipel, *ibid*, p. 37 f.

28 Seipel, *ibid.*, p. 40; Stanley Greenstein, Our Humanities Exposed – Predictive Modelling in a Legal Context, Stockholm University 2017, p. 31.

29 Tuomas Pöysti, *ICT and Legal Principles: Sources and Paradigm of Information Law*, Scandinavian Studies of Law (2004) vol. 47, p. 560.

30 Field, *supra* note 25, p. 10 f.

stunted growth in children.[31] Research regarding extreme forms of isolation such as solitary confinement in prisons have also shown that this can cause severe psychological distress in a variety of ways such as hallucinations, increased risk of suicide, self-harm etc.[32]

When it comes to care robots, we also need to deal with the human tendency towards so-called anthropomorphism. It is a long-held idea which researchers in social psychology eventually found empirical evidence for, i.e., that humans need other humans in daily life for a wide variety of reasons. The need for human contact is so strong that it is deemed to make us prone to sometimes create humans out of non-human agents, such as pets or machines, through a process of anthropomorphism.[33]

A typical example of this is that we tend to act aggressively towards technical devices such as computers if they fail to function or 'cooperate' the way we expect, sometimes cursing at a computer or hitting it. The tendency to anthropomorphize varies between individuals, depending on age, culture or the situation at hand.[34]

Anthropomorphism could also play a role in an assessment of the possible harm that an individual might experience due to involuntary loneliness, as it is plausible that our ability to anthropomorphize could mean that we would be socially stimulated enough by the company of robots to lead a happy life. Product anthropomorphism is asserted by some to be able to help increase consumer well-being,[35] but can also be perceived as deceptive.[36] At the same time, the

31 *Ibid.*, chs. 4 and 5; Shanley Pierce, *Touch starvation is a consequence of COVID-19's physical distancing,* TMC News, (May 15, 2020), https://www.tmc.edu/news/2020/05/touch-starvation/.

32 Craig Haney, *Psychological Effects of Solitary Confinement: A Systematic Critique*, Crime and Justice (2018) vol. 47, p. 371 f. Haney argues that solitary confinement ought to be eliminated entirely for some groups and greatly reduced for others.

33 Nicholas Epley et al., *When We Need a Human: Motivational Determinants of Anthropomorphism*, Social Cognition (2008) vol. 26, p. 143 f.

34 *Ibid.*, p. 144–146.

35 Fangyuan Chen, Jaideep Sengupt & Rashmi Adaval, *Does Endowing a Product with Life Make One Feel More Alive? The Effects of Product Anthropomorphism on Consumer Vitality*, Journal of the Association for Consumer Research (2018) vol. 3, pp. 503–513.

36 Nicholas Epley, *A Mind Like Mine: The Expectionally Ordinary Underpinnings of Anthropomorphism*, Journal of the Association for Consumer Research (2018) vol. 3, p. 495.

anthropomorphizing of a robot could also be assessed as a symptom of chronic social isolation or disconnection, which has been perceived in for example individuals living in isolation with pets.[37]

It has been argued that anthropomorphism and the process of dehumanization, meaning that individuals fail to attribute human features to humans and think of them as non-human agents, are two sides of the same coin.[38] Dehumanization is also a risk factor that has to be considered, which was noted earlier with reference to studies concerning the social robot PARO. This seems to be the case especially when it comes to distant monitoring of care recipients by humans, something that could take place via care robots.

In a more distant future, we might also have to redefine human contact in situations where it might be less simple to distinguish humans from machines, for example in the case of development of post-human hybrids.[39]

Another issue concerns social isolation, and even more specifically, digital isolation. Information and Communication Technology (ICT) interventions, such as social media, video calls and conferences, gaming etc., can be important tools for reducing social isolation, which is especially common in certain older age groups. One conclusion is that ICT solutions for tackling social isolation appear promising in general, though there seems to be a need for more evidence in the field. However, no matter how promising ICT interventions for reducing social isolation might seem, the digital divide in society among certain generations continues to pose a problem, so-called digital isolation.[40]

From this limited overview, one conclusion is that creating a legal notion as the basis of a potential right to human contact can be challenging, since it must be a multidimensional notion and the

37 Epley et al., *supra* note 33, p. 147.

38 Epley, *supra* note 36, p. 496–597; Epley et al., *supra* note 33, p. 153.

39 Mark Kingwell, *Do Sentient AIs Have Rights? If So, What Kind?* in AI & Fundamental Rights (Claes Granmar, Katarina Fast Lappalainen & Christine Storr, eds.) Publit 2019, p. 35–54.

40 Yi-Ru Regina Chen & Peter J. Schulz, *The Effect of Information and Communication Interventions on Reducing Social Isolation in the Elderly: A Systematic Review*, Journal of Medical Internet Research (2016) vol. 18, https://www.jmir.org/2016/1/e18/PDF.

interpretation of such a notion is certain to vary from case to case.[41] Nevertheless, this attempt at defining human contact and its implications for our well-being and, in the long run, the social stability of our society, will be used as a starting point for the legal assessment concerning the meaning of human dignity and the prohibition of inhumane and degrading treatment as laid down in Art. 3 of the ECHR and Art. 4 of the EU Charter. These are the constitutional foundation for how and to what extent care robots can be used in public care as well as for the supervision of such usage by government agencies.

3.3 Human dignity

The mother of all human rights, as some might claim, human dignity is a core principle, known as a common constitutional tradition in national legal systems within the EU, as well as in European and international law.[42] This is the case in particular regarding healthcare and care in general, where human dignity is central.[43] It is also a crucial concept when it comes to human-centric AI.[44]

Human dignity is a complicated concept.[45] Beyond the idea that human beings should not be perceived as simple objects, it can be explained as 'a bundle of more concrete conditions that must be met in order to safeguard human dignity',[46] or as 'differentiated dignity'[47]. It can also be distinguished as having two different aspects, the

41 Brownlee (2013), p. 213.

42 Sebastian Heselhaus & Ralph Hemsley, *Human Dignity and the European Convention on Human Rights*, in Handbook of Human Dignity in Europe (Paolo Becchi & Klaus Mathis eds.), Springer, Cham. https://doi-org.ezp.sub.su.se/10.1007/978-3-319-27830-8_47-1, p. 3; Barak, *supra* note 24, p. 157, referring to the German concept of *Muttergrundrecht*.

43 In for example Swedish healthcare and social service legislation, the principles of human dignity and self-autonomy are explicitly regulated.

44 Greenstein, Stanley, *Predictive Modelling, Scoring and Human Dignity*, in AI & Fundamental Rights, Granmar, Claes, Fast Lappalainen, Katarina and Storr, Christine (eds.), Publit, (2019), p. 122; AI-HLEG (2019) p. 12.

45 Greenstein, *ibid.,* (2019), p. 118.

46 Alexy, *supra* note 24, p. 233.

47 Antoine Buyse, *The Role of Human Dignity in ECHR Case-Law*, ECHR Blog (October 21, 2016), https://www.echrblog.com/2016/10/the-role-of-human-dignity-in-echr-case.html;

first is that every human life has an intrinsic value, and the second concerns the personal responsibility for one's own life.[48] In view of the definitional difficulties, it is easier to understand human dignity through its interpretation in case law.

Despite being depicted as a cornerstone of our fundamental rights, human dignity can be regarded as a 'latecomer to human rights'.[49] Article 1 of the Universal Declaration of Human Rights of 1948 reads that 'All human beings are born free and equal in dignity and rights'. This provision is not legally binding, although it can be the object of certain report procedures.[50]

Human dignity is not explicitly regulated in the ECHR, probably due to the fact that the drafters had a more practical approach in mind with the goal of creating a practice-oriented instrument, liberated from 'solemn and emphatic language';[51] rights ought to be 'practical and effective, not theoretical and illusory', as is often stated by the ECtHR.[52] Nevertheless, it plays an important part in the jurisprudence of the ECtHR, especially in regard to Art. 3 on the prohibition of torture and inhuman and degrading treatment as well as the right to privacy in Art. 8, where direct reference has been made to human dignity in several leading cases.[53]

Human dignity is also part of EU constitutional law. This right was first developed in the case law of the Court of Justice of the European Union (CJEU) but nowadays holds a prominent position in Art. 1 of the EUCFR and as part of the values of the EU laid down

48 Ronald Dworkin, Is Democracy Possible Here – Principles for a New Political Debate, Princeton University Press 2008, p. 20 f.

49 Heselhaus & Hemsley, *supra* note 42, p. 4.

50 *Ibid.,* p. 6.

51 Buyse, *supra* note 47; Costa, Jean-Paul, *Human Dignity in the Jurisprudence of the European Court of Human Rights* in Understanding Human Dignity (Christopher MacCrudden, ed.), Oxford University Press 2013, p. 394.

52 See, for example, *Mamatkulov and Askarov v. Turkey*, appl. n° 46827/99, 46951/99, Judgment (Grand Chamber) 4 February 2005, para 121.

53 *Selmouni v. France*, application nr 25803/94, Judgment (Grand Chamber) 28 July 1999, para 99. The Court stated that: '…in respect of a person deprived of his liberty, recourse to physical force which has not been made strictly necessary by his own conduct diminishes human dignity and is in principle an infringement of the right set forth in Art. 3'; *Pretty v. The United Kingdom*, application n° 2346/02, Judgment 29 April 2002, para 52 (Art. 3).

in Art. 2 of the Treaty of the European Union.[54] In the explanatory notes on Art. 1 of the EUCFR, it is clearly stated that 'none of the rights laid down in this Charter may be used to harm the dignity of another human person', which means that the right to human dignity has to be respected even where this means that another right is restricted.[55] It is therefore essential to the interpretation of other rights and can be used to extend the scope of a human right, such as is the case of the prohibition of inhuman and degrading treatment, or even to establish unwritten rights.[56] Under this line of reasoning, a right to human contact could possibly be found on the basis of the principle of human dignity.

If we assume there is a right to care by humans, a fundamental question becomes to what extent care given by robots, where an individual receives little or no physical contact from human caregivers, and where this individual does not have any other possibilities for human contact from any other human beings, such as a partner, child, or friend, would be intruding on the human dignity of that individual.

In order to examine this, it is not enough to make an analysis based on human dignity, but it is also necessary to examine if such a situation might amount to inhuman or degrading treatment, as stipulated in Art. 3 of the ECHR and Art. 4 EUCFR.

3.4 Prohibition of torture and inhuman and degrading treatment

Caregiving is supposed to be about taking care of people in a respectful and compassionate manner. However, this is a delicate matter and can involve individuals in particularly vulnerable situations, who can easily be subject to both intentional and unintentional abuse or ill treatment. Furthermore, budgetary restrictions, staff shortages and poorly organised caregivers can lead to ill treatment and neglect of

54 See, for example, Case C-377/98, *Netherlands v. Parliament and Council*, Judgement of 9 October 2001, p. 77 (concerning legal protection of biotechnological inventions).

55 *Explanations relating to the Charter of Fundamental Rights (2007/C 303/02), Explanation on Article 1 – Human Dignity*, Official Journal of the European Union, https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32007X1214(01)&from=EN.

56 Heselhaus & Hemsley, *supra* note 42, p. 3; Dworkin, *supra* note 48, p. 37 (on legal rights and political rights).

certain vulnerable groups in society.[57] It is therefore no surprise that the prohibition of torture and inhuman and degrading treatment is relevant to caregiving. The prohibition of torture and inhuman and degrading treatment in European Human Rights Law also has a history as a response to the Nazis' use of torture during World War II, which included both harmful and painful medical experiments, such as those conducted by the infamous Dr Mengele.[58]

The right to protection against torture and inhuman and degrading treatment is absolute, which means that no derogations are permitted, not even in the event of a public emergency threatening the nation.[59] The prohibition of torture and inhuman and degrading treatment is closely intertwined with that of human dignity, which has been used to extend the scope of this absolute right. This makes Art. 3 of the ECHR somewhat different from other absolute rights that are often interpreted narrowly, to the detriment of that which it is supposed to protect.[60]

The prohibitions laid down in Art. 3 of the ECHR are applicable in situations where the state is responsible for the care of individuals, such as in hospitals, prisons, or mental care facilities. These situations make it especially relevant to assess the risks related to robotization of caregiving. The protection against such treatment embraces both physical and mental pain or suffering and enshrines, as stated by Nowak and Charbord, a 'right to physical and spiritual integrity'.[61] This is regarded as a fundamental value of a democratic society.[62]

The equivalent provision in Art. 4 of the EUCFR has the same wording as Art. 3 of the ECHR and, by virtue of Art. 52 (3), the same

57 See, for example, William B.T. Mock, *Human Rights and Aging*, Generations: Journal of the American Society on Aging (2019–2020) vol. 43, pp. 80–86.

58 Nigel S. Rodley, *Integrity of the Person* in International Human Rights Law 2nd edition (Daniel Moeckli et al., eds.), Oxford University Press 2014, p. 175.

59 *Ireland v. U.K.*, application n° 5310/71, Judgment 18 January 1978, para 163.

60 Heselhaus & Hemsley, *supra* note 42, p. 14.

61 Manfred Nowak & Anne Charbord, *Art. 4 – Prohibition of Torture* in The EU Charter of Fundamental Rights – A Commentary (Steve Peers et al., eds.), Hart Publishing 2014, p. 74.

62 *Soering v. U.K.* appl. n° 14038/88, Judgment 7 July 1989, para 88 (extradition).

scope.[63] In contrast to what is seen from the CJEU, there is an abundance of case law from the ECtHR concerning Art. 3 of the ECHR. Thus, the following account will focus solely on the case law of the ECtHR.[64] What is of particular interest is the case law that involves the effects of isolation and especially solitary confinement, where the individual is, to some degree, cut off from the outside world.

Differentiating between torture, inhuman or degrading treatment is a matter of degrees in relation to the intensity of the interference.[65] On one end of the spectrum, torture constitutes an intention to inflict severe pain or suffering, whether corporal or mental, for a specific purpose, on an individual who is in a state of helplessness.[66]

The Court includes certain key variables in its assessment, such as physical and mental effects, duration, age, state of health, sex and vulnerability. It also pays attention to the nature of the context at hand as well as if use of force can be justified.[67]

When it comes to issues regarding physical or social isolation, strict forms of solitary confinement can be part of an overall assessment of what amounts to 'torture', as is shown in the case of *Ilascu* and others. In this case an individual was detained for eight years in strict isolation in a remote and dilapidated location, where he had no contact with other prisoners, was not allowed to receive and send mail and thus had no news from the outside world. Furthermore, he

63 *Explanations relating to the Charter of Fundamental Rights (2007/C 303/02), Explanation on Article 4 – Prohibition of torture, inhuman and degrading treatment or punishment*, Official Journal of the European Union, https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32007X1214(01)&from=EN.

64 The EUCFR is only applicable when union law is to be applied according to 51(1) of the EUCFR, which normally do not include for example care of the elderly or the disabled. Nowak and Charbord point out that the specialised bodies set up by the EU in the areas of security, justice and freedom, such as Frontex, EASO and Europol, are especially exposed to liability under Art. 4 EUCFR: Nowak and Charbord, *supra* note 61, p. 64.

65 Natasa Mavronicola, Torture, Inhumanity and Degradation under Article 3 of the ECHR: Absolute Rights and Absolute Wrongs, Hart Publishing 2021, p. 90; Heselhaus and Hemsley, *supra* note 42, p. 14.

66 Nowak & Charbord, *supra* note 61, p. 81.

67 Mavronicola, *supra* note 65, p. 94–112.

did not have the right to contact his lawyer or receive regular visits from his family.[68]

The ECtHR stated as a general principle that:

> […] prohibition of contact with other prisoners for security, disciplinary or protective reasons does not in itself amount to inhuman treatment or punishment. On the other hand, complete sensory isolation, coupled with total social isolation can destroy the personality and constitutes a form of inhuman treatment which cannot be justified by the requirements of security or any other reason.[69]

Another assessment was made in a case regarding the notorious terrorist known as Carlos the Jackal, who was similarly detained in solitary confinement for a total of eight years following his conviction. He was prohibited to have contact with other prisoners, but had access to TV, newspapers and was allowed regular visits from his many lawyers, priests, and family. He was able to exercise 2–3 hours daily and had visits from a doctor twice a week. The Court held that his isolation was relative and that the physical conditions in detention were satisfactory.[70]

In *Ahmad and others*, the Court had to assess whether extradition to the U.S. with the risk of incarceration at the ADX Florence prison, which is one of the most restrictive prison regimes in the country, would be contrary to Art. 3. The Court concluded that:

> Although inmates are confined to their cells for the vast majority of the time, a great deal of in-cell stimulation is provided through television and radio channels, frequent newspapers, books, hobby and craft items and educational programming. The range of activities and services provided goes beyond what is provided in many prisons in Europe.[71]

68 *Ilascu and others v. Moldova and Russia*, appl. n° 48787/99, Judgment 8 July 2004, para 438 and 440.

69 *Ibid.,* para 432 with reference to the decision in Messina N° 2 case, appl. n° 25498/94, decision 8 June 1999.

70 *Ramirez Sanchez v.* France, appl. n° 59450/00, Judgment 4 July 2006, para. 131–135.

71 *Babar Ahmad and others v. the U.K*, appl. n° 24027/07, 11949/08, 36742/08, para 222.

Keeping in mind the long-term effects of isolation that can be severely detrimental, it can nevertheless be determined to be necessary to keep someone under another regime than the ordinary one, based on the danger they pose.[72] Nonetheless, the ECtHR has repeatedly stated that procedural safeguards must be in place to guarantee the prisoner's welfare and the proportionality of the measure 'in order to avoid any risk of arbitrariness resulting from a decision to place a prisoner in solitary confinement.'[73]

Inhuman treatment, on the other hand, does not have to meet all of these criteria and does not have to be the result of sadistic intentions. It can be triggered in situations where the infliction of pain has been humiliating. Degrading treatment can be at hand if the infliction of pain or suffering is made in a particularly humiliating manner.[74] The European Committee for the Prevention of Torture and Inhuman and Degrading Treatment (CPT) has in its report regarding the Transdniestrian prison situation, which was illustrated in the *Ilascu* case, made clear its opinion that solitary confinement can, in certain circumstances, amount to inhuman and degrading treatment and that in any event solitary confinement for as many years as Ilascu and his fellow prisoners endured, was 'indefensible'.[75]

In the case of *Harakchiev and Tolumov* the ECtHR held that automatic segregation of life prisoners from the rest of the prison community and from each other, especially in combination with the lack of comprehensive activities either outside or inside the cell, might in itself raise an issue under Art. 3. The Court also pointed out that the 2006 European Prison rule 25.2, although not legally binding, clearly states that all prisoners should be allowed to spend as many hours a day outside their cells as are necessary for *an adequate level of human and social interaction*.[76]

When it comes to disabled prisoners or prisoners with mental illnesses, isolation can be deemed to be particularly aggravating. The ECtHR found that the handcuffing, in solitary confinement for

72 *Ramirez Sanchez v. France*, para. 150.

73 *Babar Ahmad and others v. the U.K*, para 212.

74 Nowak & Charbord, *supra* note 61, p. 81.

75 *Ilascu and others v. Moldova and Russia*, para 289.

76 *Harakchiev and Tolumov v. Bulgaria*, appl. n° 15018/11, 61199/12, Judgment 8 July 2014, para 204.

seven days out of nine, of a man suffering from chronic schizophrenia amounted to inhuman and degrading treatment.[77] In another case, *Z.H.* a deaf, dumb, mentally disabled as well as illiterate man unable to use the official sign language, charged with mugging, was held in prison for almost three months. The Court especially considered the circumstances regarding:

> […] the inevitable feeling of isolation and helplessness flowing from the applicant's disabilities, coupled with the presumable lack of comprehension of his own situation and of that of the prison order, must have caused the applicant to experience anguish and inferiority attaining the threshold of inhuman and degrading treatment, especially in the face of the fact that he had been severed from the only person (his mother) with whom he could effectively communicate.[78]

Living conditions in psychiatric institutions or social care homes must also provide for an adequate level of human and social interaction. The placement of a schizophrenic man against his will in a psychiatric institution in a remote mountain location under poor living conditions was deemed to be contrary to Art. 3 as well as Art. 5 regarding the right to liberty and security.[79]

The tragic case of *Valentin Câmpeanu* also touches upon isolation. The ill treatment of Mr Câmpeanu, an orphan, mentally disabled, infected with HIV, who had spent all of his life in different institutions, led to his death in a psychiatric ward at 18 years of age. Among other atrocities, it was reported that he had spent his last time in life:

> […] alone in an isolated, unheated and locked room, which contained only a bed without any bedding. He was dressed only in a pyjama top. At the time he could not eat or use the toilet without assistance. However, the staff at the PMH refused to help him, allegedly for fear that they would contract HIV. Consequently, the only nutrition provided to Mr Câmpeanu was glucose, through

77 *Kucheruk v. Ukraine*, appl. n° 2570/04, Judgment 6 September 2007, paras 134–146.

78 *Z.H. v. Hungary*, appl. n° 28973/11: Judgment 8 November 2012.

79 *Stanev v. Bulgaria*, appl. n° 36760/06, Judgment (Grand Chamber) 17 January 2012.

> a drip. The report concluded that the hospital had failed to provide him with the most basic treatment and care services.[80]

It is also noteworthy that the CPT, regarding the Swedish system of pre-trial detention in remand prisons, where solitary confinement is more the rule than the exception, has made statements to the effect that the system is unsatisfactory and in need of fundamental change. This has occurred ever since the CPT started its investigations of Sweden in 1991. There might be reasons for keeping remand prisoners in solitary confinement, even though it is not yet clear if they are guilty of a crime. In Sweden, however, the majority of remand prisoners (68%), including some juveniles, have been subject to some kind of restriction, with most of them spending up to 23 hours per day alone in their cells, lacking anything to occupy themselves with. This is reportedly detrimental to their psychological well-being.[81]

Situations related to expulsions of migrants can also occur, such as in the case of a 91-year-old woman with multiple health issues and no social network in her home country, or the case of a 9-year-old girl, an orphan, being sent back by the authorities without their making sure she had somewhere to go on arrival in the state she came from.[82]

However, Art. 3 does not apply in cases of minor mistreatment, such as certain kinds of chastisement for disciplinary purposes. The ECtHR does not suggest that any kind of behaviour on the part of officials that might be perceived as humiliating is either moral or appropriate, but nor does it necessarily reach the degree of severity required for it to fall within the scope of Art. 3.[83]

80 *Case of the Centre for Legal Resources on behalf of Valentin Câmpeanu v. Romania,* Judgment 17 July 2014, para 23.

81 CPT/Inf (2016) 1, *Report to the Swedish Government on the visit to Sweden carried out by the European Committee for the Prevention of Torture or Inhuman and Degrading Treatment from 18 to 28 May 2015*, paras 48–53, https://rm.coe.int/1680697f60.

82 *Chyzehvska v. Sweden*, appl. n° 60794/11, decision 25 September 2012; *Nsona v. the Netherlands*, appl. n° 23366/94, Judgment 28 November 1996.

83 *Costello-Roberts v. the United Kingdom*, application n° 13134/87, judgment 25 March 1993, para. 31 and 32. However, see partly dissenting opinions by Judge Ryssdal, Thór Vilhjálmsson, Matscher and Wildhaber. It is, however, not clear if the Costello-Roberts case would apply today considering the Grand Chamber judgement in *Bouyid v. Belgium*, appl. n° 23380/09, Judgment (Grand Chamber) 28 September 2015, which concerned slaps in the face of a remand prisoner, which was found to be covered by

3.5 Is there a right not to be left alone?

The most express protection against sensory and social isolation can be found in Art. 3 of the ECHR, which contains a certain protection against such forms of isolation in some state institutions, especially prisons. Complete sensory isolation, coupled with total social isolation, which cannot be justified by requirements of security or some other reason constitutes a form of inhuman treatment is strictly prohibited.

The ECtHR also makes its assessment regarding restrictive measures on the basis of various variables, which in the cases accounted for above include mental health and age. It indicates that solitary confinement is not appropriate or should at least be kept at a minimum for certain groups in a population, such as persons with mental illness, juveniles and the elderly.

However, most of these cases concern prisoners. The situation in these cases is therefore somewhat different from that in a more or less voluntary institution like a nursing home, care at home or care services for the elderly and disabled which take place in the individual's home environment. However, in cases where public institutions are responsible for the care of individuals, even in their own homes, it can be questioned to what extent individuals with for example dementia or grave physical disabilities can be left alone most hours of the day without any effective access to the outside world and human and social contact. Such conduct by publicly financed caregivers could in fact amount to inhuman and degrading treatment.

Concerning the use of care robots, this analysis could indicate that measures where individuals are left solely to the care of robots, or for long periods of time, can possibly constitute abusive care amounting to inhuman and degrading treatment, since it would disregard the human need for sensory and social contact.

The overarching principle or right to human dignity could also be supportive regarding the interpretation of certain legal requirements regarding caregiving, which could be used as a more fine-tuned instrument in deciding which levels of human contact are necessary or appropriate in situations that fall outside the scope of

---

Art. 3, constituting a serious attack on the individual's dignity (para 103). See Mavronicola, *supra* note 65, p. 98.

Art. 3, for example, in the case of a more large-scale introduction of care robots.[84]

Do these conclusions answer the question if there is a human right not to be left alone? Yes, to a certain degree.

4 Final remarks

A basic right to human contact and a right not to be left alone are implicit in the human rights protection of the ECHR as well as the EUCFR. They can also be based on evidence from other disciplines, such as medicine and psychology.

The situation that individuals would be left in the company of only care robots is of course possible, at least in theory, but it does not seem likely that this would develop into a widespread practice. The mere risk of such a development for certain groups of individuals in a vulnerable position is nevertheless serious enough to take action.

However, if a human right to human contact is going to permeate the legal system as a whole, it is not simply enough to refer to human rights. Undertaking specific regulatory measures regarding caregiving seems to be the most appropriate way forward. That is not to say that it is possible to regulate an ideal amount of human contact that each and every individual should have, since individuals can be very different and have different needs, but there is certainly a relative urgency for developing a regulatory framework in the field. Care robots might not be mandatory or even mainstream in caregiving today, but history shows that technological advances can happen quickly. The context in which care robots are used is complex and multifaceted. Apart from the imminent health dangers of loneliness and/or isolation, there are still many questions as to how humans will respond to an increasingly digital way of life, considering our abilities to anthropomorphize and how we respond to using ICT for social interactions in the long run.

84 Zardiashvili and Fosch-Villaronga (2020), p. 139. The authors state that '…we conclude by giving the policy advice to formulate an overarching, omnibus governance solution for robotics that will be based on the concept of human dignity.'

Consequently, there is a need for further research, analysis, planning and preparation for this inevitable development.[85] This also means requiring the relevant authorities to consider their future role in assessing and investigating caregiving in the context of care robots. There are certainly enough reasons to contemplate the words of Murty: 'The greatest challenge facing us today is how to build a people-centred life and a people-centred world.'[86]

85 *Ibid.*

86 Murthy, *supra* note 4, p. 242.

Explainable AI in the European Union: An Overview of the Current Legal Framework(s)*

MARTIN EBERS

* This work was supported by the Estonian Research Council grant no PRG124 and by the Research Project "Machine Learning and AI-Powered Public Service Delivery", RITA1/02-96-04, funded by the Estonian Government.

Explainable Artificial Intelligence (XAI) is relevant not only for developers who want to understand how their system or model works in order to debug or improve it, but also for those affected by such technology. Determining why a system arrives at a particular algorithmic decision or prediction allows us to understand the technology, develop trust for it and – if the algorithmic outcome is illegal – initiate appropriate remedies against it. Additionally, XAI enables experts (and regulators) to review decisions or predictions and verify whether legal regulatory standards have been complied with. All of these points support the notion of opening the black box. On the other hand, there are a number of (legal) arguments against full transparency of Artificial Intelligence (AI) systems, especially in the interest of protecting trade secrets, national security and privacy.

Accordingly, this paper explores whether and to what extent individuals are, under EU law, entitled to a right to explanation of automated decision-making, especially when AI systems are used.

1 Introduction

1.1 The Automated, Scored Society

Today, AI systems[1] based on machine learning (ML)[2] are widely employed to make decisions with far-reaching impacts on individuals and society. Many important decisions, which were historically made by people, are now either made by machines or at least prepared by them.[3] We live in a "scored society"[4] in which citizens, consumers and legal entities are increasingly subject to actions and decisions made by or with the assistance of AI systems. ML algorithms are used by private companies in almost all fields, including financial services, manufacturing, farming, engineering, transport, telecom, retail, travel, transport, logistics and healthcare.[5] *Governmental institutions* have also become increasingly reliant on algorithmic systems to analyze and predict behavior in order to make decisions. Tax offices now use algorithms to predict abuse and fraud in tax returns

1 There is currently no generally accepted definition of the term "AI". For an overview, see Sofia Samoili et al., *AI Watch. Defining Artificial Intelligence: Towards An Operational Definition and Taxonomy of Artificial Intelligence*, European Union Joint Research Centre Technical Report (2020), publications.jrc.ec.europa.eu/repository/bitstream/JRC118163/jrc118163_ai_watch._defining_artificial_intelligence_1.pdf; Independent High Level Expert Group on Artificial Intelligence, *A Definition of AI: Main Capabilities and Disciplines,* European Commission (2019), https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=56341; Peter Norvig & Stuart Russell, Artificial Intelligence: A Modern Approach, 3rd ed. (2011).

2 As regards different types of ML algorithms, cf. Ben Buchanan & Taylor Miller, *Machine Learning for Policymakers. What It Is and Why It Matters*, The Cyber Security Project, Harvard Kennedy School, Belfer Center for Science and International Affairs (2017); Mehryar Mohri, Afshin Rostamizadeh & Ameet Talwalkar, Foundations of Machine Learning (2018).

3 Cf. AI Now Institute, *AI Now 2019 Report* (2019), https://ainowinstitute.org/AI_Now_2019_Report.html; AlgorithmWatch, *Automating Society Report 2020* (2020), https://automatingsociety.algorithmwatch.org/wp-content/uploads/2020/10/Automating-Society-Report-2020.pdf.

4 Danielle Keats Citron & Frank A. Pasquale, *The Scored Society: Due Process for Automated Predictions,* 89 Wash. L. Rev. 1 (2014).

5 For an overview on different use cases, cf. Organization for Economic Cooperation and Development (OECD), *Artificial Intelligence In Society*, pp. 47 ff (2019), https://doi.org/10.1787/eedfee77-en; International Electrotechnical Commission (IEC), *Artificial Intelligence Across Industries,* pp. 45 ff (2018), https://www.iec.ch/basecamp/artificial-intelligence-across-industries.

and to allocate cases for human review.[6] In social welfare systems, algorithms are used to determine whether a citizen should be flagged because of an increased risk for irregularities or potential fraud.[7] In the field of public security, many agencies use AI systems to detect terrorists,[8] screen people at the border[9] and predict and respond to crime ("predictive policing").[10] In the US, algorithmic prognosis instruments are even used by courts to calculate the likelihood of an accused person committing another crime while on parole.[11]

The use of AI systems can improve the efficiency, effectiveness and fairness of decisions. It can speed up administrative procedures and decrease use of manpower and financial resources. There is also potential for improving the accuracy of decisions, as AI can enhance fact analyses, forecasts and legal application. AI applications can

6 David DeBarr & Maury Harwood, *Relational Mining for Compliance Risk*, United States Internal Revenue Service (2004), http://www.irs.gov/pub/irs-soi/04debarr.pdf.

7 In Spain (https://algorithmwatch.org/en/story/spain-legal-fight-over-an-algorithms-code/); in Austria (https://algorithmwatch.org/en/story/austrias-employment-agency-ams-rolls-out-discriminatory-algorithm/), in Sweden (https://algorithmwatch.org/en/rogue-algorithm-in-sweden-stops-welfare-payments/); in Finland (https://www.tieto.com/en/success-stories/2018/the-city-of-espoo-a-unique-experiment/); in the Netherlands (https://bijvoorbaatverdacht.nl/).

8 In the EU, the European Commission is funding the DANTE experiment, an anti-terrorism project (Detecting and analyzing terrorist-related online contents and financing activities), aimed at using automated decision-making against terrorism; https://www.h2020-dante.eu/.

9 Cf. Frontex, Artificial Intelligence-based Capabilities for the European Border and Coast Guard, Final Report (2021), https://frontex.europa.eu/assets/Publications/Research/Frontex_AI_Research_Study_2020_final_report.pdf.

10 Lindsey Barrett, *Reasonably Suspicious Algorithms: Predictive Policing at the United States Border*, 41 N.Y.U. Rev. L. & Soc. Change 327 (2017); Andrew Guthrie Ferguson, *Predictive Policing and Reasonable Suspicion*, 62 Emory L.J. 259, 317 (2012); Michael L. Rich, *Machine Learning, Automated Suspicion Algorithms, and the Fourth Amendment*, 164 U. Pa. L. Rev. 871 (2016); Jessica Saunders, Priscillia Hunt & John S. Hollywood, *Predictions Put Into Practice: A Quasi Experimental Evaluation of Chicago's Predictive Policing Pilot*, 12 J. Experimental Criminology 347 (2016).

11 Such processes are used at least once during the course of criminal proceedings in almost every US state: Anna Maria Barry-Jester, Ben Casselman & Dana Goldstein, *The New Science of Sentencing*, The Marshall Project (April 4, 2015), https://www.themarshallproject.org/2015/08/04/the-new-science-of-sentencing#.xXEp6R5rD. More than 60 predictive tools are available on the market, many of which are supplied by companies, including the widely-used COMPAS system from Northpointe.

also be helpful tools in overcoming human shortcomings – such as cognitive distortions, prejudices and contingencies that we are not always aware of – and have a positive effect on human rights, for example by accelerating and simplifying public administration or allowing for faster legal proceedings.

On the other hand, AI applications raise a wide variety of ethical and legal challenges.[12] AI systems can unpredictably harm people's life, health and property. They can also lead to breaches of fundamental rights, including the rights to human dignity and self-determination, privacy and personal data protection, freedom of expression and of assembly, non-discrimination or the right to an effective judicial remedy and a fair trial, as well as consumer protection.

1.2 The Black Box Problem

Of particular concern in relation to ML techniques is the opacity of many automated/algorithmic decision-making (ADM) systems. The notion of black box AI refers to scenarios in which we can see only input data and output data for algorithm-based systems, without having insight into exactly what happens in between.[13] Only a few types of AI systems are directly interpretable for the user, such as decision trees or linear and logistic regression. In contrast, the vast majority of AI systems (support vector machines, ensemble tree methods such as random forests, gradient boosting machines and, especially, deep neural networks) have a degree of opacity that makes it hard to understand how the algorithmic decisions or predictions of the systems have been reached.[14]

12 For an overview, cf. Martin Ebers, *Regulating AI and Robotics: Ethical and Legal Challenges*, in Algorithms and Law, pp. 37–99 (Martin Ebers & Susana Navas Navarro eds., 2020); Brent Daniel Mittelstadt et al., *The Ethics of Algorithms: Mapping the Debate*, Big Data & Society 1–21 (2016).

13 Additionally, it might be that the inputs themselves are entirely unknown or only partially known.

14 For a comparison between the varying degrees of explainability of different AI systems, see David Gunning, *Explainable Artificial Intelligence (XAI)*, DARPA, https://www.darpa.mil/attachments/XAIIndustryDay_Final.pptx; Bernhard Waltl & Roland Vogl, *Explainable Artificial Intelligence – The New Frontier in Legal Informatics*, Jusletter IT 22 (2018).

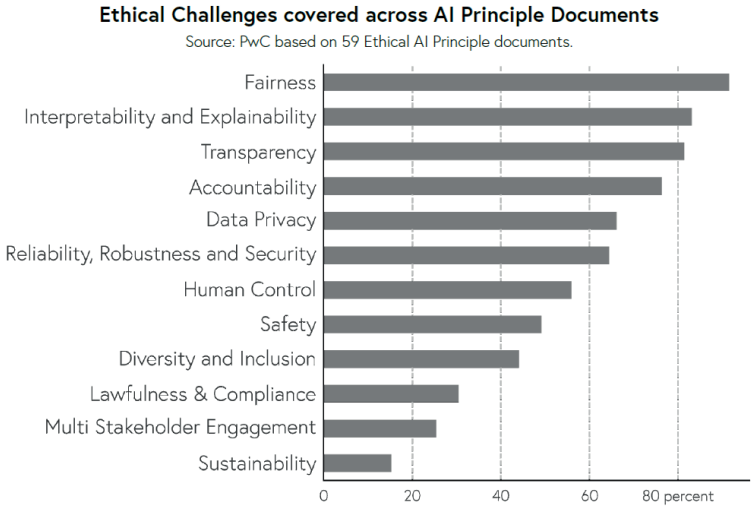


Figure 1: Ethical Challenges covered across AI Principle Documents, Source: Stanford AI Index Report 2019, p. 149.

In the Stanford AI Index Report 2019 (cf. Figure 1), interpretability, explainability and transparency of AI systems are identified (along with fairness) as the ethical challenges most frequently mentioned across 59 ethical AI principle documents.[15]

Indeed, explainability is relevant for a number of reasons.[16] For a researcher or developer, it is crucial to understand how their system or model is working in order to debug or improve it. For those affected by an algorithmic decision or prediction, it is important to comprehend why the system arrived at this conclusion in order to understand the decision, develop trust in the technology, and – if the ADM outcome is illegal – initiate appropriate remedies against it. Lastly, yet critically, explainability enables experts (and regulators) to review ADM for compliance with legal regulatory standards.

15 Raymond Perrault et al., *The AI Index 2019 Annual Report*, AI Index Steering Committee, Stanford Human-Centered AI Institute, p. 149 (2019), https://hai.stanford.edu/sites/default/files/ai_index_2019_report.pdf.

16 Avishek Anand et al., *Effects of Algorithmic Decision-Making and Interpretability on Human Behavior: Experiments using Crowdsourcing* (2018), www.l3s.de/~gadiraju/publications/HCOMP18.pdf.

In recent policy papers, the European Commission specifically highlights the latter two aspects. Thus, for example, the Commission states in its White Paper on AI[17] that the opacity of many AI systems

> "(…) may make it hard to verify compliance with, and may hamper the effective enforcement of, rules of existing EU law meant to protect fundamental rights. Enforcement authorities and affected persons might lack the means to verify how a given decision made with the involvement of AI was taken and, therefore, whether the relevant rules were respected. Individuals and legal entities may face difficulties with effective access to justice in situations where such decisions may negatively affect them."

Further, the Commission emphasizes that[18]

> "(…) there is a need to examine whether current legislation is able to address the risks of AI and can be effectively enforced, whether adaptations of the legislation are needed, or whether new legislation is needed."

1.3 Overview

In this paper, I explore the EU legal frameworks for XAI as well as the European Commission's proposal to regulate AI. To this end, the following sections deal with the question of whether EU law provides individuals with a right to explanation in the case of AI-driven decisions or predictions. Accordingly, I will consider the General Data Protection Regulation (GDPR) and the fundamental right to privacy (section B), the right to due process and fair trial (section C) and contract and tort law (section D), as the most important sources for a potential right to explanation. After this, I will look at conflicting interests that might counteract a right to explanation, especially trade secrets, intellectual property rights, national security and privacy (section E). The paper concludes with a critical analysis of the European Commission's proposal to regulate AI (section F).

The scope of the following analysis must be limited – both geographically and in terms of content. First of all, legal systems other

17 European Commission, *White Paper On Artificial Intelligence – A European Approach to Excellence and Trust*, COM/2020/65 final, p. 12 (2020).

18 *Id.*, p. 10.

than that of European Union law will not be examined in detail. While it is true that several international organizations[19] and individual countries,[20] as well as business associations and NGOs,[21] have taken actions to provide an ethical or legal framework for XAI systems, an analysis of these rules or principles would go beyond the scope of this paper.

Second, this paper will not address the question of how explainability can be achieved. In this respect, a distinction is usually made between *external explanations,* based on a set of properties used by an external observer, and *internal explanations,* based on a set of properties used by the designer (for example the source code, the parameters within the algorithms, or the weights learned by a neural network).[22] Furthermore, two broad aims of work on interpretability have been recognized in the literature: transparency and post hoc interpretability.[23] While *transparency* describes how easily a model can be understood, *post hoc interpretability* refers to how easily a

19 Cf. especially Organization for Economic Cooperation and Development (OECD), *Recommendation of the Council on Artificial Intelligence*, OECD/LEGAL/0449 (2019), https://legalinstruments.oecd.org/en/instruments/oecd-legal-0449; European Commission for the Efficiency of Justice (CEPEJ), *European Ethical Charter on the Use of Artificial Intelligence in Judicial Systems and Their Environment*, Council of Europe (2018), https://rm.coe.int/ethical-charter-en-for-publication-4-december-2018/16808f699c.

20 In France, the Digital Republic Act (Loi No 2016-1321 du 7 octobre 2016 pour une République numérique) provides that, in the case of state actors making a decision "on the basis of algorithms", individuals have a right to be informed about the "principal characteristics" of the decision-making system. For more details, see Lillian Edwards & Michael Veale, *Enslaving the Algorithm: From a 'Right to an Explanation' to a 'Right to Better Decisions'?*, 16 IEEE Security & Privacy 46 (2017).

21 For an overview of ethical initiatives in the field of AI, cf. Eleanor Bird et al., *The Ethics of Artificial Intelligence: Issues and Initiatives*, Panel for the Future of Science and Technology (STOA), European Parliamentary Research Service, pp. 37 et seq. (2020), http://www.europarl.europa.eu/RegData/etudes/STUD/2020/634452/EPRS_STU(2020)634452_EN.pdf.

22 Maja Brkan & Gregory Bonnet, *Legal and Technical Feasibility of the GDPR's Quest for Explanation of Algorithmic Decisions: of Black Boxes, White Boxes and Fata Morganas*, 11 European Journal of Risk Regulation 18–50, at 20ff (2020).

23 Bruno Lepri et al., *Fair, Transparent, and Accountable Algorithmic Decision-making Processes: The Premise, the Proposed Solutions, and the Open Challenges*, 31 Philosophy & Technology 611–627 (2018), https://link.springer.com/article/10.1007/s13347-017-0279-x; Zachary C. Lipton, *The Mythos of Model Interpretability* 16 Queue 31–57 (2018).

decision (or prediction) can be explained.[24] Approaches to post hoc interpretability can be achieved in various ways,[25] for example by visualizing what a model has learned (*visualization*), by analyzing the parameters for a single decision (*local explanations*), and by finding and presenting examples (*explanation by example*, including *counterfactual explanations*).[26] While all of these approaches are important for addressing the question of how AI-driven decisions can be explained in a comprehensible way, this paper deals solely with the preceding problem of whether individuals, under EU law, are actually entitled to explanations at all.

2 Privacy, Data Protection and Explainability

2.1 The (Missing) Right to Explanation in the GDPR

So far, most of the debate on a possible right to explanation has focused on data protection law and on the question of whether the GDPR includes such a right in the case of automated decisions.[27] The discussion centers on two provisions in particular.

2.1.1 Art. 22(3) GDPR

First, Art. 22(3) GDPR states that, in certain cases of automated processing, "the data controller shall implement suitable measures to safeguard the data subject's rights and freedoms and legitimate interests, at least the right to obtain human intervention on the part

24 Brent Mittelstadt, Chris Russell & Sandra Wachter, *Explaining Explanations in AI*, in Proceedings of the Conference on Fairness, Accountability, and Transparency, pp. 279–288 (2019).

25 Lipton, *supra* note 24.

26 Regarding counterfactual explanations, cf. esp. – in this volume – de Vries.

27 Finale Doshi-Velez, *Accountability of AI Under the Law: The Role of Explanation*, arXiv:17110 1134 (2017); Bryce Goodman & Seth Flaxman, *EU Regulations on Algorithmic Decision-Making and A Right to Explanation*, 38 AI Magazine 50–57 (2017); Gianclaudio Malgieri & Giovani Comandé, *Why a Right to Legibility of Automated Decision-Making Exists in the General Data Protection Regulation*, 7 International Data Privacy Law 243–265 (2017); Andrew D. Selbst & Julia Powles, *Meaningful Information and the Right to Explanation*, 7 International Data Privacy Law 233–242 (2017); Sandra Wachter, Brent Mittelstadt & Luciano Floridi, *Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation*, 7 International Data Privacy Law 76–99 (2017).

of the controller, to express his or her point of view and to contest the decision." In addition, recital (71) GDPR points out that these safeguards "should include specific information to the data subject," including "an explanation of the decision reached after such assessment." However, since the GDPR mentions a right to explanation only in its non-binding recitals, not in the operative text of Art. 22(3) GDPR itself, most scholars agree that this provision does not provide for a right to explanations of individual decisions.[28]

2.1.2 Art. 15(1)(h) GDPR

Second, Art. 15(1)(h) GDPR stipulates that, in the case of automated processing in the sense of Art. 22(1) GDPR, the data controller must provide data subjects with "meaningful information about the logic involved." This provision is interpreted quite differently by different legal scholars. According to one view, information can only be "meaningful" if the explanation enables the data subject to contest a decision as provided by Article 22(3) GDPR.[29] Accordingly, all information necessary to understand a decision and to check its accuracy needs to be provided.[30] Others argue, however, that Art. 15(1)(h) GDPR refers only to the general structure and functionality of an ADM system, not to the individual circumstances of a specific automated decision, and especially not to the weighing of features, machine-defined case-specific decision rules, or information about reference or profile groups.[31] In the same vein, Article 29 Data Protection Working Party (WP29) in its revised guidelines on ADM and profiling acknowledged that Art. 15(1)(h) GDPR obliges the controller to provide information "about the *envisaged consequences* of the processing, rather than an explanation of a *particular* deci-

28 Goodman & Flaxman, *id.*; Mario Martini, *Regulating Algorithms: How to Demystify the Alchemy of Code?*, in Algorithms and Law, pp. 100–135, at p. 117 (Martin Ebers & Susana Navas Navarro eds., 2020); Selbst & Powles, *supra* note 27; Wachter, Mittelstadt & Floridi, *supra* note 27.

29 Selbst & Powles, *id.*

30 *Id.*

31 Malgieri & Comandé, *supra* note 27; Wachter, Mittelstadt & Floridi, *supra* note 27.

sion"[32] and confirmed that controllers are not required to disclose the "full algorithm."[33]

In any case, both Art. 22(3) and Art. 15(1)(h) GDPR refer to ADM procedures in the sense of Art. 22(1) GDPR, which applies only to decisions based "solely" on automated processing and have "legal effects" or "similarly significantly affect" a person. Therefore, AI-based systems which only support humans in decision-making are beyond the scope of both provisions.[34] Since most algorithmically prepared decisions still involve a human being, the majority of ADM procedures are not covered by Art. 22(3) and Art. 15(1)(h) GDPR.[35]

2.1.3 Analysis

Altogether, there appears to be an overwhelming argument against deriving a right to specific explanations from the GDPR. Whether such a right exists is, of course, ultimately a matter that can only be decided by the Court of Justice of the European Union (CJEU).

At the end of the day, however, there is little ground for assuming that the CJEU would grant such a right. Legislative history challenges such a right. During the legislative drafting of the GDPR, a more ambitious "right to explanation" was discussed, but it was not implemented in the final version.[36] Moreover, CJEU case law on the

32 Article 29 Data Protection Working Party, *Guidelines on Automated Individual Decision-Making and Profiling for the Purposes of Regulation 2016/679*, European Commission, p. 27.

33 *Id.*, p. 25.

34 Benedikt Buchner, *Artikel 22 DSGVO*, in DS-GVO: Datenschutz-Grundverordnung, 2nd ed., para. 16 (Jürgen Kühling & Benedikt Buchner eds., 2018); Martini, *supra* note 28, p. 112; Wachter, Mittelstadt & Floridi, *supra* note 27, pp. 88, 92; Thomas Wischmeyer, *Artificial Intelligence and Transparency: Opening the Black Box*, in Regulating Artificial Intelligence pp. 75–101, at p. 83 (Thomas Wischmeyer & Timo Rademacher eds., 2020).

35 Wachter, Mittelstadt & Floridi, *supra* note 27, p. 92. Bygrave, on the other hand, is of the opinion that decisions formally attributed to humans but originating "from an automated data-processing operation the result of which is not actively assessed by either that person or other persons before being formalised as a decision" would fall under the category of "automated decision-making": Lee A. Bygrave, *Automated Profiling: Minding the Machine: Article 15 of the EC Data Protection Directive and Automated Profiling*, 17 Computer Law & Security Review 17 (2001).

36 Wachter, Mittelstadt & Floridi, *supra* note 27, p. 81.

former Data Protection Directive 95/46 casts doubt on such a right. The CJEU made it clear in a number of cases[37] that data protection law primarily aims at the fairness of data processing (procedural fairness) and is not intended to ensure the accuracy of decisions and the decision-making process.[38] Should the CJEU confirm this view for the GDPR, it seems unlikely to assume a right to specific explanations related to decisions that cannot be reviewed under data protection anyway.

2.2 Opaqueness of AI Systems and the Fundamental Right to Privacy

The use of opaque AI systems is not only subject to the requirements of the GDPR, but also – in the case of public bodies[39] – to the fundamental right to privacy, which is enshrined inter alia in Art. 8 of the European Convention of Human Rights (ECHR) and Art. 7 and 8 of the EU Charter of Fundamental Rights (EU-CFR).

2.2.1 Dutch Rechtbank Den Haag *in the SyRI Case*

The SyRI case,[40] decided by the Dutch *Rechtbank Den Haag* (The Hague District Court) on February 5, 2020, is the first successful example of a case of this kind, where a court ruled that an opaque algorithmic risk scoring system, the "Systeem Risico Indicatie" (SyRI), violated Art. 8 ECHR.

37 CJEU, joined cases C-141/12 and C-372/12 *YS and M and S* ECLI:EU:C:2014:2081; case C-434/16; Case C-434/16 *Peter Nowak* ECLI:EU:C:2017:994.

38 Cf. also the case law analysis by Evelien Brouwer & Frederik Zuiderveen Borgesius, *Access to Personal Data and the Right to Good Governance During Asylum Procedures After the CJEU's YS and M and S Judgment (C-141/12 and C-372/12)*, 17 European Journal of Migration and Law 259–272 (2015); Sandra Wachter & Brent Mittelstadt, *A Right to Reasonable Inferences: Re-thinking Data Protection Law in the Age of Big Data and AI*, 2 Columbia Bus. L. Rev. 494–620 at 521ff (2019).

39 The traditional scope of fundamental rights is the vertical relationship between the state and citizens. In principle, neither the ECHR nor the EU-CFR apply directly to private parties; in other words, they do not have direct horizontal effect; cf. European Union Charter of Fundamental Rights, Art. 51(1), 2007/C 303/01 (2007); opinion of Advocate General Trstenjak, delivered on 8 September 2011 in Case C-282/10 *Dominguez v Centre Informatique Du Centre Ouest Atlantique*, ECLI:EU:C:2011:559, paras 80ff.

40 *Rechtbank Den Haag*, judgment of 5.2.2020, ECLI:NL:RBDHA:2020:1878.

The SyRI system was deployed by the Dutch state to detect various forms of fraud, including social benefits, allowances and tax fraud. Based on a big data analysis of personal data, SyRI produced a risk report which indicated whether a legal or natural person could be deemed worthy of investigation with regard to possible fraud, unlawful use and non-compliance with legislation.[41] Civil rights associations criticized the use of this system because of its lack of transparency and its potential discriminatory effects on poor and vulnerable citizens.[42] In the proceedings before the Dutch court, the opacity of the system was a point of particular importance.

The SyRI system's lack of transparency played a pivotal role in the *Rechtbank Den Haag* ruling in favor of the plaintiffs (the citizen rights NGOs). According to the Court, the SyRI legislation failed to comply with Art. 8(2) ECHR because it did not strike a fair balance between the interests of the community as a whole, which the legislation served, and the rights of the individuals affected by the legislation with respect to their private lives and homes. In particular, the Court highlighted that the SyRI legislation was insufficiently transparent and verifiable as it did not provide sufficient information, especially on the functioning of the risk model – for instance, the type of algorithms used in the model – nor did it provide information on the risk analysis method as applied by the Social Affairs and Employment Inspectorate.[43] According to the Court, this lack of transparency resulted in an inability to verify how the decision tree was generated and what steps it was comprised of. Consequently, data subjects could not defend themselves against the risk report and assess whether the system produced unjustified or discriminatory results.[44]

41 *Rechtbank Den Haag*, *id.* at 3.2.

42 Gianluca Misuraca & Colin van Noordt, *Overview of the Use and Impact of AI In Public Services In the EU*, JRC Working Paper JRC 120399, p. 45 et seq. (2020), https://publications.jrc.ec.europa.eu/repository/bitstream/JRC120399/jrc120399_misuraca-ai-watch_public-services_30062020_def.pdf; Natasha Lomas, *Blackbox Welfare Fraud Detection System Breaches Human Rights, Dutch Court Rules*, TechCrunch (February 6, 2020), https://techcrunch.com/2020/02/06/blackbox-welfare-fraud-detection-system-breaches-human-rights-dutch-court-rules/.

43 *Rechtbank Den Haag*, *supra* note 40 at 6.86 and 6.89.

44 *Id.* at 6.90–6.94.

As a result, the *Rechtbank Den Haag* concluded that the SyRI legislation was in breach of Art. 8(2) ECHR, unlawful and therefore carried no binding effect.

2.2.2 Analysis

The decision of the *Rechtbank Den Haag* was enthusiastically embraced by NGOs and the UN Special Rapporteur on extreme poverty and human rights, Philip Alston.[45] Indeed, the judgment has the potential to become a landmark ruling, setting a strong legal precedent for other courts in Europe to follow. For the first time, an algorithmic system was found to be in breach of the fundamental right to privacy, mainly because of the opacity of the decision-making system.

However, it is important to note that the Dutch Court did not categorically exclude the use of ADM systems, nor did it rule that there should be full disclosure. At the end of the day, the ruling concerned only the SyRI system and its lack of legal safeguards. In its judgment, the *Rechtbank Den Haag* explicitly noted that the unlawfulness of the SyRI legislation does not mean that the State is under any obligation to disclose the risk models and risk indicators to the claimants.[46] Accordingly, it is difficult to derive from the decision a right to specific explanations of individual decisions.[47] Lastly, the significance of the judgment must also be put into perspective insofar as it concerns the use of algorithmic systems by public administrations only, not by private companies.

The decision is, nevertheless, a "wake-up call"[48] for public administrations and governments across Europe in that it makes clear that human rights laws in Europe must be central to the design and implementation of algorithmic decision-making systems.

45 UN Human Rights, *Landmark Ruling by Dutch Court Stops Government Attempts To Spy On the Poor – UN Expert*, Office of the High Commissioner (February 5, 2020), https://www.ohchr.org/en/NewsEvents/Pages/DisplayNews.aspx?NewsID=25522&LangID=E.

46 *Rechtbank Den Haag*, *supra* note 40 at 6.115.

47 Likewise, Anne Meuwese, *Regulating Algorithmic Decision-Making One Case at the Time: A Note on the Dutch 'SyRI' Judgment*, 1 European Review of Digital Administration & Law 209, 211 (2020).

48 *Id.*

3 Due Process, Fair Trial Rights and Explainability

3.1 Transparency as a Central Principle in Administrative and Judicial Proceedings

While a right to explanation cannot yet be (unambiguously) derived from existing EU secondary law or the fundamental right to privacy, such a right might follow from other constitutional safeguards recognized under the ECHR and the EU-CFR.

Transparency is an underlying prerequisite of numerous constitutional and procedural principles and rights, which emerges from the rule of law, recognized at the European level in Art. 2 of the Treaty on European Union. The rule of law requires a system of certain and foreseeable rules, where everyone has the right to be treated equally by all decision-makers, in accordance with the law, and to have the opportunity to challenge decisions through fair proceedings before independent and impartial courts.[49] The key rationale behind the rule of law lies in the promise of legal certainty.[50] It ensures that individuals can predict what they may and may not do, and also granting them knowledge about the consequences of their decisions and actions.

One of the cornerstones of the rule of law is the *right to due process* and *fair trial* – enshrined in Art. 6 ECHR and Art. 47 EU-CFR.[51] Both provisions grant the right of access to a court,[52] the right to a

49 Cf. European Commission for Democracy through Law (Venice Commission), *Rule of Law Checklist*, No. 15, Council of Europe (March 18, 2016), https://www.venice.coe.int/webforms/documents/default.aspx?pdffile=CDL-AD(2016)007-e.

50 See Friedrich A. Hayek, The Constitution of Liberty, chapters 9 and 10 (1960); Jeremy Waldron, *The Rule of Law and the Importance of Procedure*, 50 Nomos 3–31 (2011).

51 The level of legal protection required by Art. 47 CFR goes beyond the level of protection required by the ECHR. Cf. Martin Ebers, Rechte, Rechtsbehelfe und Sanktionen im Unionsprivatrecht, pp. 253 ff (2016); Oliver Dörr, Der europäisierte Rechtsschutzauftrag deutscher Gerichte, pp. 50 ff (2003); Adrienne de Moor-van Vugt, *Administrative Sanctions in EU Law*, 5 Review of European Administrative Law 5–41, at 18 ff (2012); Angela Ward, *National and EC Remedies Under the EU Treaty: Limits and the Role of the ECHR*, in The Outer Limits of European Union Law, pp. 329–361, at pp. 329 ff (Catherine Barnard & Okeoghene Odudu eds., 2009).

52 European Court of Human Rights, *Guide on Article 6 of the European Convention on Human Rights, Right to a Fair Trial (Civil Limb)*, Nos. 87ff, Council of Europe (December 31, 2020), https://www.echr.coe.int/documents/guide_art_6_eng.pdf.

public hearing[53] as well as a number of additional principles which have been recognized as being attached to these rights, such as the right to adversarial proceedings,[54] the equality of arms between the parties,[55] the right to confrontation in criminal proceedings[56] and the right to have a reasoned decision.[57]

Opaque algorithms can impair these rights in various ways, depending on how they are used and for what purpose. According to *Palmiotto*,[58] opaque algorithms impact (i) the *adversarial principle*, when a party cannot contradict the opponent's allegations; (ii) the *equality of arms principle*, when algorithms create knowledge asymmetry between parties; (iii) the *right to confrontation*, when algorithms in criminal proceedings cannot be examined by the defense; and (iv) the *right to have a reasoned decision*, when algorithms do not explain or justify how a particular decision has been reached.

The last aspect is of particular importance. Without knowing how the output of an algorithmic system has been generated, it is not possible to contest the prediction or decision. Opacity leads to a lack of means to challenge algorithm-based evidence or decision support algorithms, and consequently represents a threat to fair trial rights. In this vein, *Sir Alfred Denning* pointed out already in 1949 that "every tribunal should give a reasoned decision, just as the ordinary courts do. Herein lies the whole difference between a judicial decision and an arbitrary one. A judicial decision is based on reason and is known to be so because it is supported by reasons."[59]

From the foregoing analysis, we can conclude that fair trial rights indeed set limits on the use of opaque algorithmic systems in administrative and judicial proceedings. Whether the CJEU or the Euro-

53 *Id.*, Nos. 398ff.

54 *Id.*, Nos. 355ff.

55 *Id.*, Nos. 362ff.

56 Cf. Art. 6(3)(d) ECHR.

57 European Court of Human Rights, *supra* note 52, Nos. 386ff.

58 Francesca Palmiotto, *The Black Box on Trial: The Impact of Algorithmic Opacity on Fair Trial Rights in Criminal Proceedings*, in Algorithmic Governance and Governance of Algorithms, pp. 49–70, at p. 61 (Martin Ebers & Marta Cantero Gamito eds., 2020).

59 Alfred Denning, Freedom under the Law, pp. 91 ff (1949).

pean Court of Human Rights will adopt this view, however, remains unclear. So far, neither of the courts has yet ruled on this issue.

3.2 Supreme Court of Wisconsin in *State v. Loomis*

The situation is different in the United States. The Supreme Court of Wisconsin in *State v. Loomis*[60] addressed the question whether the right to due process was violated when a person was sentenced to prison on the basis of the well-known COMPAS risk assessment tool,[61] which is criticized above all for its lack of transparency and the risk for discrimination. As the news portal ProPublica revealed in 2016, COMPAS judged black and white prisoners differently. Among other things, it was found that the probability that black inmates were identified as high risk, but did not re-offend, was twice as high as that for white inmates. Conversely, white inmates were more likely to be classified as low risk, but later re-offend.[62]

In the *Loomis* case, Mr. Loomis was accused of being involved as a driver in a drive-by shooting. Mr. Loomis entered a guilty plea but later denied involvement, stating that he drove the car only after the incident. The circuit court convicted Mr. Loomis pursuant to his guilty plea, ruling out probation on the following basis:[63] "You're identified, through the COMPAS assessment, as an individual who is at high risk to the community. In terms of weighing the various factors, I'm ruling out probation because of the seriousness of the crime and because your history, your history on supervision, and the risk assessment tools that have been utilized, suggest that you're extremely high risk to re-offend."

Later, Mr. Loomis lodged a motion for post-conviction relief, requesting a new sentencing proceeding. To substantiate his motion, he argued that the circuit court's reference to COMPAS violated

60 *State v. Loomis*, 881 N.W. 2d 749 (Wis. 2016).

61 COMPAS is an abbreviation for "Correctional Offender Management Profiling for Alternative Sanctions".

62 See Jeff Larson et al., *How We Analyzed the COMPAS Recidivism Algorithm*, ProPublica (May 23, 2016), https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm; Jon Kleinberg et al., *Inherent Trade-Offs in the Fair Determination of Risk Scores*, Working Paper, arXiv (2016) https://arxiv.org/abs/1609.05807, 5–6.

63 *State v. Loomis*, *supra* note 60 at 755.

his constitutional right to due process.[64] In support, Mr. Loomis submitted that the circuit court's use of the COMPAS assessment infringed on both his right to an individualized sentence and his right to be sentenced on accurate information, as COMPAS reports provide data relevant only to particular groups and because the methodology used to make the reports is a trade secret.[65]

However, the Wisconsin Supreme Court upheld the decision of the circuit court because it was of the opinion that the right to due process had not been violated. Although the Wisconsin Supreme Court acknowledged that the "proprietary nature of COMPAS" prevented the disclosure of how risk scores are calculated,[66] the Wisconsin Supreme Court rejected Loomis's motion, bringing forward two arguments: First, the Wisconsin Supreme Court emphasized that since COMPAS used only publicly available data and data provided by the defendant, Mr. Loomis could have denied or explained any information that went into making the report and could have verified the accuracy of the information used in sentencing.[67] Second, the Wisconsin Supreme Court emphasized that "although the circuit court referenced the risk assessment at sentencing, the [circuit] court essentially gave it little or no weight,"[68] and that the circuit court would have imposed the exact same sentence without the COMPAS risk assessment.[69]

Although Mr. Loomis's motion was not successful, the Wisconsin Supreme Court added that judges must proceed with caution when using risk assessments based on algorithms.[70] To this end, the Wisconsin Supreme Court explained in its judgment how risk assessments should be used by trial courts.[71] Moreover, the Wisconsin Supreme Court added that risk scores may not be used "to determine whether an offender is incarcerated" or "to determine the severity of

64 *Id.* at 756.

65 *Id.* at 757.

66 *Id.* at 763ff.

67 *Id.* at 761–762.

68 *Id.* at 770.

69 *Id.* at 771.

70 *Id.* at 765.

71 See *id.* at 763–765.

the sentence."[72] Lastly, the Wisconsin Supreme Court highlighted that presentencing investigation reports using the COMPAS assessment must include particular written warnings for judges, including the warning that the proprietary nature of COMPAS prevents disclosure of how risk scores are calculated.[73]

In October 2016, Mr. Loomis filed a petition with the US Supreme Court, which was denied in June 2017.[74]

Other legal proceedings in the US were also unsuccessful.[75] In *Malenchik v. State*,[76] the Indiana Supreme Court rejected the defendant's claim that using risk assessment tests in determining a sentence is unconstitutional. The Indiana Supreme Court stated that the sentence had been based on factors other than the risk assessments, since the trial court had also relied on the defendant's prior criminal history and refusal to accept responsibility for his actions and change his behavior and had not used the algorithm's output as an independent aggravating factor.

3.3 Analysis

Neither ruling is particularly convincing from the perspectives of US constitutional law[77] or European human rights law. Although the Wisconsin Supreme Court pointed out that COMPAS was a "poor fit" for sentencing, it nevertheless accepted its risk assessment as a sentencing factor. More importantly, both the Wisconsin Supreme

72 *Id.* at 769.

73 *Id.*

74 *Loomis v. Wisconsin*, 137 S. Ct. 2290 (2017) (denying cert.).

75 For an overview, cf. also Cary Coglianese & Lavi M. Ben Dor, *AI in Adjudication and Administration*, 2118 Faculty Scholarship at Penn Law, at 12ff (2020), https://scholarship.law.upenn.edu/faculty_scholarship/2118; Virginia Foggo & John Villasenor, *Artificial Intelligence, Due Process, and Criminal Sentencing*, 2020 Michigan State L. Rev. 295–354, at 333ff (2020).

76 *Malenchik v. State*, 928 N.E.2d 564, 568 (2010).

77 Cf. especially the critique by Han-Wei Liu, Ching-Fu Lin & Yu-Jie Chen, *Beyond State v Loomis: Artificial Intelligence, Government Algorithmization and Accountability*, 27 Int'l J. L. & Info. Tech. 122–141, 130ff (2019); Note, *State v. Loomis: Wisconsin Supreme Court Requires Warning Before Use of Algorithmic Risk Assessments in Sentencing*, 130 Harv. L. Rev. 1530, 1534 (2017); Leah Wisser, *Pandora's Algorithmic Black Box: The Challenges of Using Algorithmic Risk Assessments in Sentencing*, 56 Am. Crim. L. Rev. 1811–1832 at 1813ff (2019).

Court and the Indiana Supreme Court denied that any violation of due process rights had occurred, on the grounds that the risk scores were only one factor among many that played a role in the decisions.

With this kind of thinking, both courts ignored the so-called *anchoring effects* of algorithmic systems – a concept used by behavioral psychologists to describe the common human tendency (or cognitive bias) to rely too heavily on the first piece of information offered.[78] This effect can also be proven in court proceedings. Numerous studies have demonstrated that courts which receive numerical probability data are likely to give such data undue weight because of this very anchoring effect.[79] With computer-generated numerical output, these effects are further amplified by the *automation bias*,[80] which refers to the "tendency to disregard or not search for contradictory information in light of a computer-generated solution that is accepted as correct."[81] After all, how can humans "stay in the loop" and override an algorithmic risk assessment if at the same time they have reasonable ground to believe that machines can compute faster and more precisely and can handle complexity better than humans?

How the CJEU, the European Court of Human Rights or national constitutional courts will rule on a case in which an algorithmic system has (partly) influenced the outcome of a court decision remains to be seen. One possibility is that European courts prohibit the use of ADM tools in court proceedings per se. Another possibility would be to allow the use of these systems only if they meet certain minimum requirements, especially regarding fairness, auditability and transparency.

78 See Cass Sunstein, *Hazardous Heuristics*, 70 U. Chi. L. Rev. 751, 752 (2003): "[I]n the face of uncertainty, estimates are often made from an initial value, or 'anchor,' which is then adjusted to produce a final answer."

79 Christopher Stein & Michelle Drouin, *Cognitive Bias in the Courtroom: Combating the Anchoring Effect in Criminal Sentencing* (June 23, 2017), https://ssrn.com/abstract=2991611.

80 David Lyell & Enrico Coiera, *Automation Bias and Verification Complexity: A Systematic Review*, 24 J. Am. Med. Inform. Assoc. 423–431 (2017), https://doi.org/10.1093/jamia/ocw105.

81 Mary Cummings, *Automation Bias in Intelligent Time Critical Decision Support Systems*, 2 Collection of Technical Papers Aiaa 1 st Intelligent Systems Technical Conference 557–562 (2004), https://scholars.duke.edu/individual/pub1108365; Cf. also Raja Parasuraman & Victor Riley, *Humans and Automation: Use, Misuse, Disuse, Abuse*, 39 Human Factors 230–253 (1997).

Irrespective of this, one might wonder whether the fundamental right to due process is actually a suitable means of ensuring the transparency of algorithmic systems. Constitutional courts only have the power to decide ex post in a concrete case whether the use of a specific algorithmic system violated the right to due process. They cannot establish general *binding* standards for the transparency of algorithmic systems that go beyond the individual case. What's more, since the right to due process is a defensive one, it is not possible to positively derive from this right a claim of an individual to an explanation of a specific algorithmic decision.

4 Explainability and Liability for AI

Apart from data protection and constitutional law, both contract and tort law could constrain the use of non-explainable ML models.

4.1 Human Oversight and Explainability as Standard of Care

First, it can be argued that the standard of care, applicable in both contract and tort law, requires *human oversight by professional actors* (such as doctors) when an AI system is used, as this is the only way to ensure that the system does not make any errors.[82] Arguably, in some situations, it may be possible for humans to detect false predictions and decisions even without having access to detailed information about how the AI system works. However, in most cases, an evaluation of the output of AI systems requires knowledge on the part of the user as to why the system has come to a certain conclusion. Explainability of AI systems is thus a necessary condition for users to "stay in the loop" and, if necessary, overrule its results. Hence, in order to avoid liability, professional actors may soon be legally compelled by courts to use explainable ML models.[83]

82 Phillip Hacker et al., *Explainable AI Under Contract and Tort Law: Legal Incentives and Technical Challenges*, 28 Artificial Intelligence & Law 415–439, 424 (2020).

83 *Id.*

4.2 Business Judgment Rule and Explainability

Similar issues arise with regard to the standard of care that managers of companies must employ with respect to due diligence.[84] According to the *business judgment rule*, recognized in both German and US law, managers are largely exempt from liability; however, this only applies if they have made a decision on a sufficient basis of information. Here, again, liability can only be avoided if the professional actor can stay in the loop and evaluate whether a prediction might be incorrectly positive or negative, which is only possible if the outcome of the AI system is comprehensible.

4.3 Burden of Proof and Explainability

Other liability rules might also foster XAI systems; namely, if the user/operator is generally liable – either under national or European law – for damages caused by AI systems, and laws (or courts) shift the burden of proof to the detriment of the operator.

Such a provision can be found, for example, in the European Parliament's resolution of 20 October 2020 for a "civil liability regime for artificial intelligence".[85] Art. 8 of this resolution foresees a fault-based liability for operators of AI system.[86] According to Art. 8(1), the operator of an AI system is in principle liable for any harm or damage that was caused by a physical or virtual activity, device or process driven by the AI system. However, according to Art. 8(2), the operator shall not be liable if he or she can prove that the harm or damage was caused without his or her fault, which is the case for instance if "due diligence" was observed by "selecting a suitable AI-system for the right task and skills, putting the AI-system duly into operation, monitoring the activities and maintaining the operational reliability by regularly installing all available updates." Such proof is, again, very likely to succeed only if the AI system

84 *Id.*, 426ff.

85 European Parliament, *Resolution of 20 October 2020 With Recommendations to the Commission On A Civil Liability Regime for Artificial Intelligence*, 2020/2014(INL) (2020).

86 Art. 8 of the resolution applies only to operators of low-risk AI systems. If a high-risk AI system is used, the operator is strictly liable in accordance with Art. 4 of the resolution.

is sufficiently explainable for its functionality and activities to be monitored.

4.4 Results

As a result, both contract and tort law can provide incentives to develop and use XAI systems. Admittedly, there is no direct obligation to use XAI systems, under either area of law. Nevertheless, both areas can promote the use of XAI, at least indirectly.

5 Interests Conflicting with Explainability

The above analysis illustrates that there is currently no well-established right to XAI under EU law. Insofar as such a right is called for, the question arises if there are important legal grounds which make more ambitious disclosure requirements problematic.

5.1 The Risks of Opening the Black Box

Ensuring the transparency of ML applications involves various risks and negative impacts, which could preclude more ambitious disclosure requirements.

First, a right to explanation could conflict with *intellectual property rights*[87] and the *protection of trade and business secrets*.[88] Indeed, algorithms are often intentionally kept secret for the sake of competitive advantage. Granting a third party access to such information

87 As to the question whether AI systems, the underlying code, AI databases, AI training data and/or AI outputs are protected by intellectual property rights, cf. Daniel Gervais et al., *Trends and Developments in Artificial Intelligence – Challenges to the Intellectual Property Rights Framework: Final Report*, European Commission (2020), https://www.ivir.nl/publicaties/download/Trends_and_Developments_in_Artificial_Intelligence.pdf; Susana Navas, *Creativity of Algorithms and Copyright Law*, in Algorithms and Law, pp. 221–234 (Martin Ebers & Susana Navas eds., 2020). Cf. also the European Parliament's *Resolution of 20 October 2020 on Intellectual Property Rights For the Development of Artificial Intelligence Technologies*, 2020/2015/INI (2020).

88 For the EU, cf. Directive (EU) 2016/943 of the European Parliament and of the Council of 8 June 2016 on the Protection of Undisclosed Know-how and Business Information (trade secrets) Against Their Unlawful Acquisition, Use and Disclosure, O J L 157 (June 15, 20116); Jasper Siems, *Protecting Deep Learning: Could the New EU-Trade Secrets Directive Be an Option for the Legal Protection of Artificial Neural Networks?*, in Algorithmic Governance and Governance of Algorithms, pp. 137–156 (Martin Ebers & Marta Cantero Gamito eds., 2020).

could also infringe upon the fundamental rights of the company that developed the system. In this vein, recital (63) GDPR emphasizes that data access rights "should not adversely affect the rights or freedoms of others, including trade secrets or intellectual property and in particular the copyright protecting the software."

Second, keeping an AI system opaque can also be important for *ensuring its effectiveness*, for example to prevent spambots from using the disclosed algorithm to attack the system or prevent people from cheating the system by tilting the outputs of an AI system in a desired direction.[89]

Another potentially negative effect of transparency concerns *privacy and data protection*. Making available the training or input data of the ML algorithm may violate privacy and the GDPR, if the dataset enables identification of personal data. Lastly, opacity might also be necessary to *protect national security*.[90]

5.2 Tools for Balancing Competing Interests

Given the aforementioned competing interests, far-reaching transparency requirements for private actors must be justified in light of these actors' fundamental rights. On the other hand, legislatures must also take into account the public interests at stake as well as the risky nature of AI systems. Moreover, the law also needs to protect the fundamental rights of those negatively affected by AI systems. As a result, interests counteracting transparency will hardly ever prevail. Even legitimate claims to AI secrecy do not justify blanket exceptions.[91] What is necessary, instead, is an approach based on balancing competing interests in transparency and secrecy.

In this respect, *Wischmeyer*[92] has shown that there are already various regulatory tools available which can be employed to ensure the

89 Christian Sandvig et al., *Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms*, in 64th Annual Meeting of the International Communication Association, 1, 9, (2014) http://social.cs.uiuc.edu/papers/pdfs/ICA2014-Sandvig.pdf.

90 Matthias Leese, *The New Profiling: Algorithms, Black Boxes, and the Failure of Anti-Discriminatory Safeguards in the European Union*, 45 Security Dialogue 494 (2014).

91 Wischmeyer, *supra* note 34, p. 85, with reference to recital (63), sentence 6 GDPR.

92 *Id.*, p. 85.

protection of truly sensitive data while still providing valuable information about a system and its operations.

These include:

- temporal restrictions of access rights,
- the employment of information intermediaries so that sensitive information is not given publicly, but only to trusted third parties bound to secrecy (if necessary, under criminal law protection of the obligation to maintain secrecy), who in turn prepare an anonymized summary and/or evaluate the information,
- multi-tiered access regimes for information which distinguish between different data sources, for instance by giving access to confidential data only to actors obliged to keep results secret (e.g., supervisory authorities, auditors), whereas affected persons have only limited access rights,
- procedural safeguards, for example by excluding the public in administrative and judicial proceedings.

In many cases, these tools will increase transparency without negatively affecting the secrecy interests of system operators. Thus, legitimate interests in secrecy do not speak, across the board, against a right to explanation of algorithmic decisions. Rather, what is required is to strike a balance between conflicting interests.

6 The Way Forward

The above analysis reveals that EU law does not yet provide a well-established legal basis for a right to explanation of decisions or predictions based on AI systems. Admittedly, there is sufficient evidence to recognize such a right, especially when AI systems are used in administrative or judicial proceedings. However, as long as such a right is neither explicitly acknowledged in written EU law nor derived from interpretation by the CJEU, the European Court of Human Rights or national courts, there is considerable legal uncertainty to the detriment of those who have been negatively affected by algorithmic decisions.

This final section investigates the European Commission's recent proposal for an Artificial Intelligence Act and examines how this reg-

ulation, if adopted, could contribute to the transparency of AI-based decisions.

6.1 The European Commission's Proposal for an Artificial Intelligence Act

On April 21, 2021, the European Commission presented its long-awaited proposal for a regulation laying down harmonized rules on AI, the so-called Artificial Intelligence Act.[93] The new rules would apply directly to both public and private actors inside and outside the EU, as long as their AI system is placed on the EU market or its use affects people located in the EU.[94]

The draft regulation follows a risk-based approach, which differentiates between four categories, e.g., AI systems that create (i) unacceptable risks, (ii) high risks, (iii) limited risks, and (iv) minimal risks:

- AI systems that create *unacceptable risks* due to their threat to the safety, livelihood and rights of people are banned according to the proposal.[95] This includes social scoring by governments, exploitation of vulnerabilities of specific groups of persons (e.g., children), the use of subliminal techniques, and – subject to exceptions – real-time remote biometric identification systems used in publicly accessible spaces for law enforcement.
- *High-risk AI systems* are permitted on the EU market; however, they are subject to compliance with certain mandatory requirements and an ex ante conformity assessment before they can be put on the market.[96] Annex III of the proposal contains a list of high-risk AI systems and can be updated by the European Commission.[97] For these systems, mandatory requirements apply

93 *Commission Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts*, COM/2021/206 final (2021).

94 *Id.*, Art. 2(1).

95 *Id.*, Art. 5.

96 *Id.*, Art. 8ff.

97 So far, the Commission has identified high-risk systems in eight areas, e.g., AI technology used in (i) critical infrastructures (e.g., transport); (ii) educational or vocational training (e.g., scoring of exams); (iii) safety components of products (e.g., AI application in robot-assisted surgery); (iv) employment, workers' management and

regarding the quality of datasets used; technical documentation and record-keeping; transparency and the provision of information to users; human oversight; and robustness, accuracy and cybersecurity.

- For certain AI systems with *limited risks*, the draft regulation also foresees transparency obligations to make sure that people know they are communicating with or facing an AI system.[98] This encompasses AI systems that interact with humans (e.g., chatbots), emotion recognition and biometric categorization systems, and systems that generate or manipulate content (deep fakes).
- Systems with *minimal risks*, i.e., all other AI systems, can be developed and used in conformity with already existing legislation, without any additional legal obligation. Providers of those systems may choose to voluntarily apply the requirements for trustworthy AI and adhere to voluntary codes of conduct.[99]

The explainability of AI systems is regulated in the draft regulation only for high-risk systems. According to Art. 13(1)(1) of the proposal, high-risk AI systems "shall be designed and developed in such a way to ensure that their operation is sufficiently transparent to enable users to interpret the system's output and use it appropriately." Moreover, Art. 13(2) of the proposal states that high-risk AI systems shall be "accompanied by instructions for use in an appropriate digital format or otherwise that include concise, complete, correct and clear information that is relevant, accessible and comprehensible to users." Additionally, the proposal foresees technical documentation and record-keeping requirements for high-risk AI systems.[100] Recital (47) of the proposal highlights that, in order "to address the opacity that may make certain AI systems incomprehensible […] a certain degree of transparency should be required for high-risk AI systems.

---

access to self-employment (e.g., CV sorting); (v) essential private and public services (e.g., credit scoring denying citizens opportunity to obtain a loan); (vi) law enforcement that may interfere with people's fundamental rights (e.g., evaluation of the reliability of evidence); (vii) migration, asylum and border control management (e.g., verification of authenticity of travel documents); and (viii) administration of justice and democratic processes (e.g., applying the law to a concrete set of facts).

98 Commission Proposal, *supra* note 93, Art. 52.

99 *Id.*, Art. 69.

100 *Id.*, Art. 11–12.

Users should be able to interpret the system output and use it appropriately. High-risk AI systems should therefore be accompanied by relevant documentation and instructions of use and include concise and clear information, including in relation to possible risks to fundamental rights and discrimination, where appropriate."

6.2 Critical Assessment

The requirements for explainability of high-risk AI systems laid down in Art. 13 of the proposal are certainly a step in the right direction. However, it is rather problematic that this norm only formulates general requirements for transparency, without specifying them. The proposal is silent on the specific measures that need to be taken to ensure that AI systems are sufficiently transparent. Instead, this issue is largely left to the self-assessment of the provider, who must ensure that the system undergoes an appropriate conformity assessment procedure before it is placed on the market or put into service.[101]

While it is true that a provider must demonstrate, upon request of a national competent authority, that their AI system complies with the transparency requirements set out in Art. 13,[102] the question of how to make an AI system explainable is left to the discretion of the AI system provider.

Additionally, the draft places a great deal of trust in harmonized standards developed by private standardization bodies. According to the proposal, standardization "should play a key role to provide technical solutions to providers to ensure compliance with this Regulation."[103] Hence, AI systems which are in conformity with harmonized standards[104] shall be presumed to be in conformity with

101 *Id.*, Art. 16(a) and (e).

102 Cf. *id.*, Art. 16(j) and Art. 23.

103 *Id.*, Recital (61).

104 International, European and national standardization organizations are in the process of developing such technical standards for AI systems. For an overview of the existing standards that apply to AI systems and the ongoing standardization work in the field of AI cf. International Electrotechnical Commission, *Artificial Intelligence Across Industries,* pp. 71 ff (2018), https://www.iec.ch/basecamp/artificial-intelligence-across-industries; DIN & DKE, *German Standardization Roadmap on Artificial Intelligence*, pp. 147 et seq. and pp. 155 et seq. (2020), https://www.din.de/resource/blob/772610/e96c34dd6b12900ea75b460538805349/normungsroadmap-en-data.pdf.

the mandatory requirements of the regulation, to the extent that the standards encompass those requirements.[105] Against this backdrop, some critical voices have already highlighted that many AI applications are at risk of being "whitewashed" by this compliance mechanism.[106]

Another problem is that neither providers nor users are required, under the proposal, to provide transparency to the person affected by an AI-based prediction or decision. According to Art. 13, the provider must ensure transparency only vis-à-vis the user of the system.

Moreover – and this is arguably one of the most crucial points – the draft regulation does not provide for any individual rights. Although the regulation is intended to protect fundamental rights, it lacks remedies by which individuals can seek redress for any breach of the regulation. In particular, the draft does not foresee any mechanism to facilitate individuals' recourse against AI-driven decision-making.

Accordingly, the draft regulation does not provide for a right of affected persons to receive an explanation of algorithmic decisions, nor does it provide for a right to inspect the internal documentation of the ML model or at least the key decision factors.

6.3 Outlook

Although the European Commission's proposal for an Artificial Intelligence Act is to be welcomed, there is considerable room for improvement. Since a right to explain the outcome of algorithmic systems cannot be clearly derived from existing EU law, such a right – together with access rights to documentation – should be clearly established in the proposed AI regulation,[107] at least for high-risk

105 Commission Proposal, *supra* note 93, Art. 40.

106 Yannick Meneceur, *Proposition de règlement de l'IA de la Commission européenne: Entre le trop et le trop peu?*, Les Temps Électriques (April 4, 2021), https://lestempselectriques.net/index.php/2021/04/22/proposition-de-reglement-de-lia-de-la-commission-europeenne-entre-le-trop-et-le-trop-peu/.

107 It is sometimes proposed to include such a right in the GDPR. However, this is problematic for two reasons. First, there is also a need to establish access rights for non-personal data; second, these access rights should be given to auditors other than data protection authorities; cf. Philipp Hacker, *AI Regulation in Europe*, p. 10 (2020), https://ssrn.com/abstract=3556532.

systems. Otherwise, persons affected by AI-driven decisions will not be able to determine if their rights have been respected.

Such rights would not necessarily imply an obligation of AI providers to open the black box and reveal trade secrets. Rather, in order to protect the legitimate business interests of companies, access rights to internal documentation should be granted only to actors obliged to keep results secret (e.g., supervisory authorities, auditors), whereas affected persons should have at least a right to receive an explanation of algorithmic decisions and a right to access a list of key decision factors.

Additionally, the rights to explanation and access to documentation could be complemented by a new liability regime with rebuttable presumptions and reversals of the burden of proof. For example, the burden of proof for damages caused by AI systems could be linked to compliance with the obligations foreseen in Art. 11 and 12 of the proposed AI regulation regarding technical documentation and record-keeping. Hence, a breach of these duties could trigger the rebuttable presumption that a particular damage was caused by a defective AI system.[108] Such rules would not only contribute to the overall transparency of AI systems and strengthen the incentives for careful AI development, but could also help solve some of the most pressing problems that currently exist in the area of private enforcement for those affected by AI-driven decisions.

108 Such a rule is also suggested in the Commission's White Paper on AI and the Liability Report; cf. European Commission, *White Paper On Artificial Intelligence – A European Approach to Excellence and Trust*, COM/2020/65 final, p. 15 (2020); European Commission, *Report on the Safety and Liability Implications of Artificial Intelligence, the Internet of Things and Robotics*, COM/2020/64 final, p. 15 (2020).

Algodicy: Justifying Algorithmic Suffering

Can Counterfactual Explanations be used for Individual Empowerment of those Subjected to Algorithmic Decision-Making (ADM)?*

KATJA DE VRIES

Abstract

This text introduces the term "algodicy" as a neologism to describe the justification of suffering caused by algorithms. Algodicies often operate at a statistical or population level, lacking justification with regard to a particular individual. This tendency is strengthened when algorithmic decision-making (ADM) is employed in public-sector bureaucratic practices dominated by a style of governing where steering the population as a whole takes precedence over impact on individual cases. The text discusses the UK grading controversy (summer 2020) to exemplify how algorithmic suffering is (not) justified at an individual level. The UK grading algorithm, which is an example of a relatively simple equation based on top-down hypotheses articulated by its human creators, is contrasted with more complex models generated by machine learning (ML) techniques such as neural networks. Here, the justification of the ADM is further complicated by a lack of transparency and interpretability, making reliance on (too) high-level algodicies even more attractive. As an alternative to algodicies, the use of counterfactual explanations is explored. Counterfactual explanations operate at the level of the individual, by providing the nearest hypothetical example that would have resulted in a different ADM classification. Such a hypothetical "*What if…?*" can act as a way to justify algo-

* Sections IV and V of this paper contain some parts that have been published in K. De Vries, "Transparent dreams (are made of this): Counterfactuals as transparency tools in ADM," Critical Analysis of Law, vol. 8, no. 1, 2021.

rithmic suffering at an individual level but also as an empowering tool for individuals to reverse the negative impact of ADM on their lives – showing what changes could result in a different ADM classification. Counterfactual explanations are promising tools, not least as a way to give teeth to legal rights for justifications or explanations of ADM decisions following from data protection, administrative law and legislation regulating artificial intelligence. However, for a variety of reasons – including the so-called *Rashomon effect*, which entails that there is a multiplicity of equally fitting models – counterfactual explanations are no panacea against all algorithmic suffering, and the individual might ultimately be confronted with the fundamental opacity of an algorithm that is not fully interpretable. While counterfactual explanations can be empowering for affected individuals, they are prosthetic constructions that will often be built on top of algorithmic-bureaucratic decision-making systems that are not inherently engaged with individual concerns.

1 Algodicy – a neologism to describe the justification of suffering caused by algorithms

Why is there evil, if God is good? In 1710, the philosopher Leibniz coined the neologism *theodicy* in his eponymous book to answer this question: a justification (*dikē*) of God (*theo*). According to Leibniz, notwithstanding evil, suffering and pain, we live in the best of all possible worlds. It is the lack of a bird's eye perspective that prohibits us from seeing that, despite the presence of suffering and pain,

> "…that there is an infinitude of possible worlds among which God must needs have chosen the best, since he does nothing without acting in accordance with supreme reason. Some adversary not being able to answer this argument will perchance answer the conclusion by a counter-argument, saying that the world could have been without sin and without sufferings; but I deny that then it would have been better. For it must be known that all things are connected in each one of the possible worlds: the universe, whatever it may be, is all of one piece, like an ocean: the least movement extends its effect there to any distance whatsoever, even though this effect become less perceptible in proportion to the distance."[1]

1 Leibniz, G. W. (2007). *Theodicy. Essays on the Goodness of God, the Freedom of Man and the Origin of Evil*. BiblioBazaar.

In 1983, in a modern and secularized world, the philosopher Sloterdijk reimagined Leibniz's theodicy by coining another neologism: *algodicy*, that is, a justification of pain (*algos*) in the absence of God.[2] In the face of "immeasurable suffering,"[3] it can be unbearable to admit that everything has been "for nothing."[4] Algodicy means that a secular higher meaning is used to justify suffering. Such a higher meaning can be, for example, political (e.g., "*the soldier sacrificed his life for the Fatherland*"), biological (e.g., "*we have been overexploiting Nature and now Nature strikes back*") or historical (e.g., "*the suffering of our forefathers has resulted in our current emancipation*"). The focus is always shifted from the individual level to a justification at a higher level of abstraction.

While the etymology of *algodicy* (derived from ancient Greek) is unconnected to *algorithm* (a Latinized version of the name of Persian mathematician al-Khwarizmi), the word algodicy has always popped into my mind as a fitting existential term for the human suffering following from algorithmic misclassifications and the call for transparency and explainability. In this text, I give a new meaning to Sloterdijk's neologism: I let *algodicy* refer to justifications (*dikē*) for the pain (*algos*) caused by *algo*rithms.

2 An example of ADM in the public sector: the UK algorithmic grading controversy

In the summer of 2020, in the midst of COVID-19 pandemic, the UK had a grading controversy that is an excellent example of the kind of suffering that can be caused by algorithmic decision-making (ADM). Following the lockdown in the spring of 2020, UK pupils in secondary education had been unable to take the A level[5] exams that give entry to higher education. Without exam grades, higher educational institutions would have difficulties deciding who

2 Sloterdijk, P. (1988). *Critique of cynical reason*. University of Minnesota Press.

3 *Idem*, p. 460.

4 *Idem*, p. 461.

5 In the UK, A levels (Advanced Level qualifications) are exams taken mainly by 18-year-olds to gain entry to higher education institutions. The exams are graded on a scale from A*–E, where A* is the highest grade and E is the lowest. Pupils who do not fulfil the minimum standards in their exams receive the grade U (unclassified).

to admit. One option would have been to simply give all pupils teacher-predicted grades ("center-assessed grades" or CAGs). After all, teachers probably have the greatest insight into the capabilities of their own pupils. The problem, however, is that teachers tend to be too optimistic about their pupils, and that reliance on CAGs would therefore lead to grade inflation.

> "Within society, grades are a form of currency, and their use relies on there being stability over time (unless there is a clear reason why results might change). If standards are not maintained, then the value and credibility of grades is likely to be undermined. This would be problematic given that there is a reliance within other parts of the education system (and more widely) on qualification grades being comparable over time."[6]

This meant that there was a political incentive[7] to find another solution to prevent grade inflation. In June 2020, the Office of Qualifications and Examinations Regulation (Ofqual) offered a workaround by creating a grading model[8] to predict what grade students mostly likely would have received if they had taken the exams. In this grading model – the *Direct Centre Performance model* (DCP) – several variables were used to calculate P_{kj}, a predicted distribution of grades for each individual school or college: $P_{kj} = (1 - r_j)C_{kj} + r_j(C_{kj} + q_{kj} - p_{kj})$.[9]

6 Ofqual (2020). *Awarding GCSE, AS, A level, advanced extension awards and extended project qualifcations in summer 2020: interim report*, p. 24. https://www.gov.uk/government/publications/awarding-gcse-as-a-levels-in-summer-2020-interim-report.

7 Clarke, L. (2020). Ofqual advisor: Prioritising grade inflation was a political decision. *New Statesman Tech*. https://www.newstatesman.com/spotlight/2020/08/ofqual-advisor-prioritising-grade-inflation-was-a-political-decision.

8 I use the word "model" interchangeably with the words "algorithm" and "equation". In public discourse, the word "algorithm" was used most frequently.

9 The variable $_k$ stands for a specific grade. The variable $_j$ stands for the specific school. An explanation of the algorithm can be found in Hern, A. (2020). Ofqual's A-level algorithm: why did it fail to make the grade? *The Guardian*. https://www.theguardian.com/education/2020/aug/21/ofqual-exams-algorithm-why-did-it-fail-make-grade-a-levels.

The DCP model[10], combined with a relative ranking of pupils (each teacher ranked every pupil in their class from best to weakest in a given subject), was used to adjust the CAGs provided by teachers. Individual predicted grades were fitted into the predicted grade distribution (P_{kj}). This means that if a teacher ranked a student as the top student of a class and predicted that he/she would receive the highest grade (A*), a predicted grade distribution for the school ranging from U to B would mean that the A* was downgraded to a B, as A* was outside the school's predicted grade distribution. However, if a teacher at a school that had been doing very well in a subject in the last three years (predicted grade distribution A–A*) had a very weak class, giving the top pupil in the class a B, the DCP model would upgrade that grade to an A*.

A closer inspection of the DCP model identifies three elements as relevant in predicting grade distribution: how well earlier cohorts (those graduated in 2017–2019) had performed on their A levels at *the pupil's school*, how well *the pupil's class* had performed on the GCSE[11] exams taken a few years earlier, and the expected national grade distribution given the results for that subject at *a national level* during previous years. In the DCP model, the first two elements were the preferred input variables, but if historical data for the school (C_{kj}) or earlier GSCE results for the class (q_{kj}) were lacking or not deemed to be representative, the fallback option was the national grade distribution (the third relevant variable). For example, this would mean that if a subject was taught at a particular school for the first time in 2020 and historical data were lacking, the grades in a class would be distributed based on the overall national distribution. This situation arose at Eton, a very high-achieving school, resulting in a substantial downgrading of CAGs. The head of Eton wrote to the Government:

> "Rather than accept our CAGs and/or consider alternative historic data in the previous syllabus we had been following (…), the board chose instead to take the global spread of results for 2019 and apply that to our cohort, (…). This failed to take any account of the fact

10 The DCP model was used to predict both A level grades and GSCE grades (see footnote 11). However, for the sake of simplicity, this article is focused on how the DCP model was used to predict A level grades.

11 GCSE (General Certificate of Secondary Education) are compulsory exams taken mainly by 16-year-olds.

> Eton is an academically selective school with a much narrower ability range than the global spread. The results awarded to many boys in this subject bore no relation at all to their CAGs or to their ability. Several have lost university places as a result."[12]

In the DCP model, the variable r_j stands for "how many pupils in the class actually have historical data available."[13] The availability of historical data (r_j) thus determines to what extent the predicted grade distribution (P_{kj}) depends on the prior attained results at the school (C_{kj}, that is, "the historical grade distribution at the school over the last three years, 2017–2019") and of the class on the GSCEs a few years earlier (q_{kj}). Moreover, the impact of variable q_{kj} can be tempered by the variable p_{kj}, the "predicted grade distribution of the previous years," showing if this is a school where grades attained a few years earlier on GSCEs tend to be good predictors of A-levels. In a school where earlier exams (GSCEs) do not tend to have a high correlation with results on A levels, the impact of these earlier exams is minimized and the algorithmically predicted grade of a pupil is either based on the national grade distribution or, in the case of classes with less than 15 pupils, merely on the CAG. The underlying reasoning for the latter is that statistical generalizations were deemed to be unreliable for very small school classes. Thus, CAGs were used instead of the grade prediction algorithm because statistical generalizations were deemed to be unreliable in very small school classes. These unaltered CAGs were considered to be the reason that the total percentage of very high grades (A and A*) was 2.4% higher in 2020 than in 2019 (27.6 per cent of grades in 2020 compared to 25.2 per cent in 2019), despite the fact that the grades of 39% of all students had been downgraded.[14] From a policy perspective, the overall

12 Hussain, D. (2020). 'Great relief' from school chiefs over government A-level grade u-turn after ministers heeded calls from Eton College headmaster to dump the 'unfair' algorithm. *Daily Mail*. https://www.dailymail.co.uk/news/article-8635439/Eton-College-headmaster-leads-calls-government-scrap-unfair-level-algorithm.html.

13 Hern (2020), see above, footnote 9.

14 Ofqual (2020), p. 7. See above, footnote 6. Also see: Whittaker, F. (2020) A-level results 2020: Top grades up by 2.4 percentage points. *FE Week*. https://feweek.co.uk/2020/08/13/a-level-results-2020-top-grades-up-by-2-4-percentage-points/. It should be noted that these figures refer to the situation *before* the Government was forced to take a U-turn (mid-August 2020), relinquish the Ofqual grading model and fully rely on CAGs. This resulted in substantial grade inflation. Most places at universities were already filled at

result of this workaround looked satisfying: the goal to prevent grade inflation was achieved and *overall* the grade statistics looked pretty similar to those of the previous years (not surprising, given that the goal of the grade algorithm was to mimic them). Individual discontent, Ofqual argued, is something that can be expected:

> "We know that, just as in any year, some students will be disappointed with their results. Some students may think that, had they taken their exams, they would have achieved higher grades. We will never know."[15]

The problem is that exams are supposed to be an *individual opportunity* to excel. However, the Ofqual algorithm was extending the historical status quo and looking at averages, as most algorithms do. This meant that the algorithm resulted in unfair grading predictions for so-called "black swans": the predictions would be too pessimistic about exceptionally strong students at weak schools and too optimistic about exceptionally weak students at strong schools. Moreover, because smaller classes are usually found at wealthier elite schools, the fact that CAGs (which tend to be too optimistic) were relied upon when a class contained less than 15 students added to a bias against students attending weaker and less well-funded schools.

So, was the algorithm biased against students attending weaker schools, thus disadvantaging those from a lower socio-economic background? The most obvious answer is that this was indeed the case: the CAGs that were downgraded primarily affected students attending weaker schools, while students taking A-levels at wealthier schools with smaller classes (especially in subjects that attract limited number of students such as "ancient history, Latin and philosophy"[16]) benefitted from receiving only CAGs and being excluded from potential algorithmic downgrading. However, in a 180 page

---

the time of the U-turn, creating a dilemma right before the start of the academic year 2021–2022 regarding how much flexibility was required from the universities – should they create extra spots beyond their regular capacity?. Weale, S. (2020). U-turn on exams may create new set of problems in England. *The Guardian*. https://www.theguardian.com/education/2020/aug/17/u-turn-exams-may-create-new-set-problems-england.

15 Ofqual (2020), p. 8. See above, footnote 6.

16 Amoore, L. (2020). Why 'Ditch the algorithm' is the future of political protest. *The Guardian*. https://www.theguardian.com/commentisfree/2020/aug/19/ditch-the-algorithm-generation-students-a-levels-politics.

report published in late November 2020 Ofqual dismisses such allegations as anecdotic and goes to great lengths to show that the algorithmically calculated grades were not systematically biased "against candidates with protected characteristics or from disadvantaged backgrounds."[17] Moreover, giving all students CAGs as a substitute for their A levels (as happened in mid-August, after outrage about the Ofqual algorithm increased) also results in unfairness: because CAGs are overly optimistic, a much higher proportion of students would be awarded with high grades than in previous years, resulting in inflation of the value of these grades.[18] At a general statistical level, the grades predicted by the algorithm largely followed the grade distribution of earlier years across ethnicity, gender, socio-economic background, etc., as Ofqual fiercely argued in the aforementioned reports released in the summer and autumn of 2020.[19] In fact, broadly speaking, "students from disadvantaged backgrounds were on course to do slightly better in 2020 than they had in 2019."[20] The problem here is that such a statistical algodicy does not address the unfairness at an individual level. A brilliant, hardworking student based at a weak school who gets downgraded and whose future plans are crushed based on the prediction of an algorithm will hardly be consoled by the fact that at an *overall* level the same number of students from educationally and socio-economically weak backgrounds were awarded high grades as in previous years. What seems to make

17 Lee, M.W, Stringer, N. & Zanini, N. (2020) *Student-level equalities analyses for GCSE and A level* (Ofqual report 20/6713), p. 6. https://www.gov.uk/government/publications/student-level-equalities-analyses-for-gcse-and-a-level.

18 Weale, S. (2020). See above, footnote 14. *The Guardian*.

19 Ofqual (2020); Lee, M.W, Stringer, N. & Zanini, N. (2020). See above, footnotes 15 and 17.

20 Lamont, T. (2021). The student and the algorithm: how the exam results fiasco threatened one pupil's future. *The Guardian*. https://www.theguardian.com/education/2021/feb/18/the-student-and-the-algorithm-how-the-exam-results-fiasco-threatened-one-pupils-future.

sense at a societal and statistical level[21] is deeply absurd and baselessly unfair at an individual level.[22]

One of the big problems of ADM is that its functioning is black-boxed, i.e., lacks transparency and interpretability.[23] When a decision has a negative impact on an individual's life – whether it is a bad grade, a loan application being rejected, or an administrative fine – an explanation of the underlying reasons is of utmost importance for its legitimacy and acceptability. Still, it should be underlined that transparency and interpretability do not automatically result in the legitimacy and acceptability of a decision. The underlying logic of the Ofqual grading algorithm is, at least since the moment it was publicly shared, both transparent and interpretable. In contrast to many other ADM systems, the grading algorithm is not based on any artificial intelligence (AI) or machine learning (ML). In fact, it would be more correct to speak of a grading *equation*, to avoid the misconception that this is an inductive, bottom-up, AI- or ML-generated rule whose complexity makes it opaque for human understanding. The Ofqual grading algorithm is a human-made rule that, once one understands what the different variables are, is surprisingly straightforward. It is also, from a governmental perspective, justifiable: it solves the problem of grade inflation and prevents top-ranking educational institutions from being flooded with many more applications than they can handle. Moreover, it should be noted that the overly optimistic grading of CAGs also has a strong bias – teacher-predicted grades are higher for pupils with parents who have graduate degrees.[24] There is no doubt that the problem that the

21 Research suggests that teacher assessments are as reliable and stable as standardized test scores such as A levels and GSCEs. Rimfeld, K., Malanchini, M., Hannigan, L. J., Dale, P. S., Allen, R., Hart, S. A., & Plomin, R. (2019). Teacher assessments during compulsory education are as reliable, stable and heritable as standardized test scores. *Journal of Child Psychology and Psychiatry, 60*(12), 1278-1288. This means that the fear of grade inflation might in fact be a fetishization of the status quo of "the system."

22 Lamont, T. (2021). See above, footnote 20.

23 Pasquale, F. (2015). *The Black Box Society: The Secret Algorithms That Control Money and Information*. Harvard University Press; Waltl, B., & Vogl, R. (2018a). Explainable Artificial Intelligence: The New Frontier in Legal Informatics. *Jusletter IT, 4*, 1-10; Waltl, B., & Vogl, R. (2018b). Increasing Transparency in Algorithmic- Decision-Making with Explainable AI. *Datenschutz und Datensicherheit – DuD, 42*(10), 613–617.

24 Adams, R. (2021). Teachers face 'almost impossible task' awarding A-level and GCSE grades. Study finds students in England from graduate households received

Ofqual algorithm aimed to solve is real.[25] The difficulty lies in that the solution offered by the algorithm overlooked the fact that, in a constitutional democracy like the UK, one cannot simply manage a problem like grade inflation without taking into account the rights of affected individuals – including their right to contest a decision based on an individual assessment. In a way, the logic underlying the grading algorithm resembles the Freudian logic of displacement that gives a surreal twist to dreams and jokes:

> "[A] blacksmith […] committed a capital crime. The court decided that the penalty for the crime must be paid, but since he was the only blacksmith in the village and therefore indispensable, while there were three tailors, one of the latter was hung in his stead."[26]

A justification that works from a governmental-managerial perspective can be unacceptable from the perspective of individual rights. Transparency and interpretability in themselves will not help to turn a societally and legally unacceptable logic into an acceptable one.

3 "Why me, Lord?" Explaining algorithmic decisions to those who suffer because of them

Few things are as hard to swallow as suffering that is undeserved and that cannot be altered by individual actions. A psychological experiment from the late sixties showed that dogs that received shocks, regardless of where they jumped or what they did, learned that nothing they did mattered and would sink into a passive state of "learned helplessness" resembling what we call "depression" in humans.[27] In the biblical book of Job, the main protagonist is looking for expla-

more generously assessed grades. *The Guardian*. https://www.theguardian.com/education/2021/jun/08/teachers-face-almost-impossible-task-awarding-a-level-and-gcse-grades.

25 However, see footnote 21 above. The question of how to distribute spots at universities in a fair way based on merits is a real and complex one. However, scores on A levels might not be an ideal distributive tool to begin with.

26 Freud, S. (1905). *Wit and Its Relation to the Unconscious*. Kegan Paul; Freud, S. (1920). *A general introduction to psychoanalysis* (G. S. Hall, Trans.). Boni and Liveright.

27 Seligman, M., & Maier, S. (1967). Failure to Escape. *Journal of Experimental Psychology, 74*, 1–9.

nations for his suffering. Why is he – a good, kind, righteous and god-fearing man – suddenly victim of an avalanche of pain and loss of everything that is dear to him? What Job does not know is that he is a rather randomly chosen subject in a wager between Satan and God, to see how much suffering it takes to turn a god-fearing man into someone who curses God.

> "If I have sinned, what have I done to you, you who see everything we do? Why have you made me your target? Have I become a burden to you? Why do you not pardon my offenses and forgive my sins?"[28]

When we are subjected to a decision with a negative impact – we fail to pass an exam, our asylum application is rejected, or we are denied social security benefits – we are, like Job, not looking for just any explanation, but one that is individual and actionable.[29] Job wants to know what actions he has to take to be relieved of further economic, psychological and bodily misery. Full transparency ("You're a rather randomly chosen subject in a wager") would hardly be satisfying, helpful or actionable at an individual level. However, this is not a concern, as no form of transparency is given to Job – instead he is told to respect the unfathomable mystery of God's decisions:

> "Can you fathom the mysteries of God? Can you probe the limits of the Almighty? They are higher than the heavens above—what can you do? They are deeper than the depths below—what can you know? Their measure is longer than the earth and wider than the sea."[30]

As Esposito[31] has pointed out repeatedly in her work, algorithmic rationality fulfills a similar role as the pre-modern divination prac-

28 7:20–21, The Book of Job (New international version translation). https://www.biblegateway.com/passage/?search=Job%201&version=NIV.

29 Barocas, S., Selbst, A. D., & Raghavan, M. (2020). The hidden assumptions behind counterfactual explanations and principal reasons. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, pp. 80–89.

30 11:7–9 The Book of Job (New international Version translation), see above, footnote 28.

31 Esposito, E. (2013). Digital Prophecies and Web Intelligence. In M. Hildebrandt & K. De Vries (Eds.), *Privacy, Due Process and the Computational Turn. The Philosophy of Law Meets the Philosophy of Technology* (pp. 121–142). Routledge; Esposito, E.

tices that tried to capture unfathomable, divine rationality. The inner workings of algorithms are often black-boxed and shrouded in mystery, data scientists are like priests who tell the future based on indices in the present, algorithms are said to contain informational patterns that exceed human understanding in their complexity. Justifications of the rationality of God (theodicies) and algorithms (algodicies) often follow a similar format, stating that full transparency is impossible due to the limitations of human understanding and that the justification does not operate at the level of individuals, but serves some higher cause such as the stability of the state, the divine course of history, etc. The UK grading algorithm is atypical in the sense that it is a man-made, relatively simple equation that is comprehensible. This can be contrasted to, for example, ADM based on facial recognition: if you are prevented from entering a soccer stadium because a smart camera has classified you as a banned hooligan or as looking overly agitated, the underlying reasons for this algorithmic match will often be too complex for human understanding. It should be underlined that black-boxing is not something that relates merely to the complex nature of certain algorithms. It can also follow from a conscious decision by its developers or users to preserve a trade secret in order to have a competitive edge or to prevent individuals from being able to adapt their behavior to avoid algorithmic classification (gaming the system). In the case of the UK grading algorithm, it took a while before full transparency was provided, following extensive public pressure. In terms of justification, the UK grading algorithm seems[32] to follow the typical form of the algodicy: the goal of the algorithm was to preserve the overall statistical distribution of grades from previous years and prevent grade inflation. The individual cases that would suffer from this were seen

(2018). *Future and uncertainty in the digital society* Federal Agency for Civic Education (bpb), Humboldt Institute for Internet and Society (HIIG), Berlin. https://www.bpb.de/mediathek/266822/elena-esposito-future-and-uncertainty-in-the-digital-society; Kahn, R. (2018). Omens and algorithms: A response to Elena Esposito. *Digital Society Blog. Making sense of our connected world.* https://www.hiig.de/en/omens-algorithms-response-elena-espositos-future-uncertainty-digital-society/.

32 Prevention of grade inflation has been identified as the main goal of the grading algorithm. Clarke, L. (2020), see above, footnote 7. Further, the grading algorithm was also perceived as a fair way to grade during the pandemic, by balancing out potentially inflated scores given by some teachers. Still, the prevention of grade inflation seems to have been the primary and leading goal. Ofqual (2020), see above, footnote 6.

as collateral damage that could probably be adjusted at some later stage through individual appeals.

Individual concerns can easily be crushed in *any* state bureaucracy. State bureaucracies govern and manage populations by exercising their power at the individual level, steering individual bodies and minds, but are not concerned with the individuals as individuals[33] – which means that individual fundamental rights have an important role in the vertical relationship between citizen and state, as an antidote to this population management perspective. One example can be seen in the childcare allowance scandal in the Netherlands,[34] where the tax office wrongly accused tens of thousands of parents of fraud. Following a political call to fight fraud at any cost, the tax office and a group of overzealous bureaucrats executed the harsh legislation[35] that followed therefrom, without leniency. In one extreme case (that is nevertheless an excellent example of the spirit in which the legislation was executed by the tax office and upheld by the courts), this resulted in a claim by the tax office for restitution of 18,000 euros in received child care benefits after a parent had failed to make a 190 euro payment.[36] The human suffering resulting from the child allowance scandal was so extreme – poor families pushed into an abyss of financial despair – that the Dutch government was forced to step down in early 2021. The fact that a substantive portion of the affected families had dual citizenship only deepened the scandal. Suspicions of automated ethnic profiling, which were never convincingly proven, arose when it became clear that the tax office possessed and consulted a database with information about the nationalities of citizens. Most likely, the bias arose by chance or due to human bias. Still, the robotic nature of the execution and the presence and consultation of a database containing information that

33 Foucault, M., Senellart, M., & Davidson, A. I. (2007). *Security, territory, population: lectures at the Collège de France, 1977-78*. Palgrave Macmillan.

34 Frederik, J. (2021). *Zo hadden we het niet bedoeld. De tragedie achter de toeslagenaffaire*. De Correspondent.

35 General Law on Income-Dependent Schemes: Wet van 23 juni 2005 tot harmonisatie van inkomensafhankelijke regelingen (Algemene wet inkomensafhankelijke regelingen).

36 Frederik, (2021), see above, footnote 34; Raad van State (Council of State) (2018). Judgment passed on 17 January 2018, case nr. 201703951/1/A2. Online available at: https://uitspraken.rechtspraak.nl/inziendocument?id=ECLI:NL:RVS:2018:137.

could be used for ethnic profiling, led many – including myself[37] – to falsely believe, that an element of ADM bias was involved. In fact, it turned out that this scandal was based on human bias, not algorithmic bias.[38] The confusion is not surprising: at their worst, state bureaucratic decision-making and ADM share a propensity for being cold decision-making machineries, missing the individual human element, being prone to institutionalized biases and lacking in individual accountability.

Hence, it is not surprising that when bureaucratic and algorithmic practices are conflated – which happened in the UK grading scandal (see above, section II) – this can result in a reinforcement of bad tendencies. When mixing bureaucratic practices with algorithms, which due to their statistical nature have difficulty taking the individual into account, there is a risk that one ends up with a double blind spot for individual concerns and challenges. If the algorithms are not simple human-made rules (as was the case in the UK grading controversy), but complex ML-generated models, the problem of opaqueness and lack of interpretability is added to the mix. As this makes it hard for an individual to challenge a decision, this is a third element that can make the bureaucratic use of ADM unattractive for individuals subjected to the negative decisions emanating therefrom.

To make it possible for individuals to challenge opaque algorithmic-bureaucratic practices, some form of antidote is needed to allow for that which does not come naturally. Individual rights, such as the right to respect for private life in Art. 8 of the European Convention for Human Rights, the right to obtain "meaningful information about the logic involved [in], as well as the significance and the envisaged consequences" of automated decision-making in Art. 15(1–h), and the right not to be subjected to automated individual decision-making in Art. 22(1) General Data Protection Regulation

37 De Vries, K. (2020). AI policy in the Netherlands: More focus on practice than principles when it comes to trustworthiness. In S. Larsson, C. Ingram Bogusz, & J. Andersson Schwarz (Eds.), *Human-Centred AI in the EU: Trustworthiness as a strategic priority in the European Member States* (pp. 132–157). European Liberal Forum asbl.

38 Blauw, S. (2020). Algorithms are biased, but so are people. We need to decide which bias we prefer. *The Correspondent*. https://thecorrespondent.com/288/algorithms-are-biased-but-so-are-people-we-need-to-decide-which-bias-we-prefer/3809118297 6-664bcebe.

2016/679 (GDPR), are such antidotes, but they are insufficient. In order for individual rights to be effective weapons, there must be *individual* transparency about the decisions against which they are aimed. An individual must know why a negative decision was made in his/her case, in order to challenge that decision. But how does one bring back a decision to the individual level, if the logic operates at a level above the individual? In the worst-case scenario, the bureaucratic-algorithmic logic acts above the individual level in a triple way. First, the bureaucratic goals can operate at the level of population management (prevent grade inflation). Second, the algorithm can operate by generalizing patterns in time (for example, "If you did poorly in your GSCEs two years ago, you probably will do poorly in your A levels too") and within groups (for example, "If you are in a class or school with pupils who tend to have low grades, you are likely to get weak grades too"). Lastly, the algorithm, at least when it is a complex ML-generated model, can go beyond individual understanding and have so many (interacting) variables that it becomes difficult to pinpoint exactly which variables constitute "the difference that makes a difference"[39] in an individual case. This final aspect requires further discussion (see below, section IV), before we can turn to how counterfactual explanations could help the individual when facing algorithmic suffering following from ADM (see below, section V).

4 How do counterfactual explanations compare to other types of transparency in the case of complex ML-generated models?

During the 2010s, the increasing presence of opaque ADM systems has been accompanied by a growing understanding that transparency and accountability need to be created with regard to such systems.[40] The legal and policy focus on transparency has led to an

39 Bateson, G. (1972). *Steps to An Ecology of Mind: Collected Essays in Anthropology, Psychiatry, Evolution, and Epistemology*. Jason Aronson Inc.

40 European Parliamentary Research Service. (2019). *A governance framework for algorithmic accountability and transparency.* https://www.europarl.europa.eu/RegData/etudes/STUD/2019/624262/EPRS_STU(2019)624262_EN.pdf.

extensive amount of research in this field.[41] In the case of a human-made, top-down algorithm, like that in the UK grading controversy, transparency is still a relatively straightforward matter of disclosure. However, if one uses a bottom-up, inductive ML method, the classification rule (that is, the algorithmic "model") will often be too complex to grasp. As an example, we can imagine that instead of departing from human assumptions (such as "the majority of individual grades can be predicted pretty accurately based on earlier grades obtained in the school or class"), we train a ML-grading model on labelled input data from all UK pupils who have taken A-levels in 2017–2019. For each pupil, we create a very rich dataset, which not only contains the A level grades, but also all the grades this pupil, his/her classmates and other pupils at the same school received in the last ten years, their postal code, gender, weight, height and hobbies. The ML model based in these data could become extremely complex. One could imagine such a model as a grading rule that is so complex, that the "*if…, then…*" sentence fills a book of 200 pages, and encompasses endless clauses, exceptions, lists of requirements, referrals, negations, interactive conditions, etc. The grading rule might begin by saying that *if* a pupil played soccer in 2nd grade, lives in postal code X, and pupils in the class below have a grade average of Y, unless the pupil has received grade Z in 3rd grade, etc., etc., *then* the predicted grade will be A. However, in practice, any such model has more in common with pinball than with printed text. Different individuals will trigger different parts of the grading rule and retracing the exact steps of the decision path will often be impossible. Now, if we have a grading ADM that is created by ML, there are at least four different types of transparency techniques one could consider.

First, there are transparency techniques that target the training data (following the principle "garbage in, garbage out").[42] Such tech-

41 Abdul, A., Vermeulen, J., Wang, D., Lim, B. Y., & Kankanhalli, M. (2018). Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda. Proceedings of the 2018 CHI conference on human factors in computing systems, Montreal, Canada.

42 This approach goes well with block chain technology. See, e.g., Nassar, M., Salah, K., ur Rehman, M. H., & Svetinovic, D. (2020). Blockchain for explainable and trustworthy artificial intelligence. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 10*(1), e1340.

niques can be helpful in terms of systemic transparency: for example, a civil rights NGO might highlight that the training data are based to 80% on male pupils from northern England, and that applying them to female pupils in the south might lead to poor results. However, for an individual student who wants to know why she got a D instead of an A and how she can do better on the next assessment, such transparency will be of little use.

Second, there are proponents of showing the so-called *real* machinery of the ADM black box: the source code. However, the source code will often be of little use in terms of individual action because it is too complex and uninterpretable, particularly for those who are not skilled in computer science. Moreover, the creators and owners of ADM systems might wish to avoid revealing source code to protect trade secrets or to prevent people from gaming the system.

Third, there is a set of techniques for building simpler models on top of complex models[43], to capture the essence of the latter. However, simplification always comes with the risk that something essential might get lost. Using ML to summarize more complex ML models also creates a further reliance on – to use a horrible anthropomorphism – the good *judgment* of ML-based model extraction. Moreover, this method still aims for a *global* description of the model, which is likely to be unattractive from the perspective of the owner or creator because it would be *too* revealing (cf. the aforementioned concerns about trade secrets and gaming the system), as well as from an individual perspective. If a complex grading ADM, with thousands of parameters, is simplified into a model with only the twenty most important parameters, the simplified version might give a gist of the type of parameters that are considered important in the model. However, it is entirely possible that none of the top twenty parameters was decisive in an individual case. Consequently, a simplified model does not necessarily tell a student why a particular grade was suggested in his/her *individual* case.

This is where the fourth category comes in: local explanations aiming to clarify the circumstances that led up to a particular outcome in an individual case. Counterfactual explanations are a form

43 Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *arXiv preprint nr. 1602.04938*. http://arxiv.org/abs/1602.04938.

of local explanations first proposed by Wachter et al.[44] in 2018 and have since received significant attention.[45] Wachter et al. define a counterfactual explanation as "a statement of how the world would have to be different for a desirable outcome to occur"[46], and give the following example:

> "You were denied a loan because your annual income was £30,000. If your income had been £45,000, you would have been offered a loan."[47]

Counterfactual explanations can be helpful where the transparency or interpretability of a classification falls short. For example, imagine that a pupil that has been unable to take A levels in biology due to a lockdown. Instead, the grade that the student would most likely have gotten is predicted with the hypothetical ML-generated model described earlier in this text. The pupil gets an E as a predicted grade. This pupil can get insight into the reasons for this classification (*counterfactual explanation*) by being provided with a hypothetical nearest alter ego (*counterfactual example*) which would be classified by the ML system as deserving of a D as a predicted grade. A comparison with such a synthetic lookalike or nearest neighbor could be helpful because the way in which the real pupil differs *most* from this nearest neighbor "constitutes the most likely reason for the decision."[48] Instead of engaging in the (often impossible) task of retracing every single element in an ADM-process leading up to a certain output in an individual case, one looks for "the difference

44 Wachter, S., Mittelstadt, B., & Russell, C. (2018). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology, 31*(2), 841-887.

45 Barocas, S., Selbst, A. D., & Raghavan, M. (2020), see above, footnote 29; Dandl, S., & Molnar, C. (2019). Counterfactual Explanations. In C. Molnar (Ed.), *Interpretable machine learning. A Guide for Making Black Box Models Explainable*. https://christophm.github.io/interpretable-ml-book/.

46 Wachter et al. (2018), p. 844. See above, footnote 44.

47 *Ibid.*

48 Hernandez, J. (2018). Making AI Interpretable with Generative Adversarial Networks. *Medium*. https://medium.com/square-corner-blog/making-ai-interpretable-with-generative-adversarial-networks-766abc953edf.

that makes a difference".[49] Framed in this way, counterfactual examples hold the promise of being extremely powerful tools for individual (counter-)action in ADM (mis-)classification. Are they indeed as powerful as some believe?

5 The Rashomon effect: how to pick the best counterfactual from a multiplicity of hypothetical answers to the "Why me?" question

Anyone who has ever tried their hand in the art of divination – reading the future from a palm, tea leaves or cards – will know that there are multiple ways of (mis)reading the same data. Multiple interpretations can fit the same omens equally well and there is no simple tool to decide which interpretation is the most accurate. No wonder the tragic misreading of omens is a theme in ancient Greek and Roman mythology.[50] While there are many similarities between divinatory and algorithmic logic,[51] the most important in this context is what in ML circles is called the "Rashomon effect,"[52] which entails that:

> "… different models, all of them equally good, may give different pictures of the relation between the predictor and response variables. The question of which one most accurately reflects the data is difficult to resolve. (…) with data having more than a small number of dimensions, there will be a large number of models whose fit is acceptable. There is no way, among the yes-no methods for gauging fit, of determining which is the better model."[53]

When reading the future from a palm, tea leaves or cards, there are many different ways to connect the dots and construct a narrative about the future. When you have a complex input dataset (such as

49 Bateson (1972), see above, footnote 39.

50 Raphals, L. (2013). *Divination and prediction in early China and ancient Greece*. Cambridge University Press, p. 285 ff.

51 Esposito (2013); Esposito (2018). See above, footnote 31.

52 Breiman, L. (2001). Statistical Modeling: The Two Cultures. *Statistical Science, 16*(3), pp. 199–231.

53 *Idem,* pp. 203–204.

prior attainment of graduating classes at a school, the grades of a class at an earlier exam, height, postal code, hobbies, etc.), there are different predictive models that can be created to get output data (an exam grade). One can create many different equations (combining different variables in different ways) that are all reasonably successful in predicting an exam grade. These can be distinguished in terms of their predictive accuracy. However, if you have models with similar predictive accuracy, it becomes very hard to decide which one is the best. Say that we have 20 different predictive models that all have the same level of predictive accuracy. Each of these models uses 250 input variables to predict an exam grade. Some models will use the same 250 input variables but in different ways, while others use other input variables. What criterion do we have to decide which of these models is best? In principle: none.

> "The problem is that each one tells a different story about which variables are important."[54]

This is what is called the Rashomon effect:

> "(This effect is named after) a wonderful Japanese movie in which four people, from different vantage points, witness an incident in which one person dies and another is supposedly raped. When they come to testify in court, they all report the same facts, but their stories of what happened are very different."[55]

In contrast to the Leibniz's universe, where God always picks the best of all possible worlds, there is no guarantee that humans pick the best possible model to make data tell their story. There is no simple criterion to distinguish which model is best if several models fit the data equally well. In the case of counterfactual explanations, there can be an interesting doubling of the Rashomon effect. When an ADM is constructed, it is possible that the choice of model is rather arbitrary one between equally good alternatives. Why is reality captured in this particular predictive model and not in another? Then, after the ADM makes a classification having a negative impact on the life of an individual, it is likely that a multiplicity of

54 *Idem,* p. 206.

55 *Ibid.*

counterfactual examples can be constructed: various equally fitting hypothetical nearest neighbors that each results in a different ADM classification. The counterfactuals can even be contradictory:

> "Each counterfactual tells a different 'story' of how a certain outcome was reached. One counterfactual might say to change feature A, the other counterfactual might say to leave A the same but change feature B, which is a contradiction."[56]

The crucial question here is which counterfactual to pick. Dandl and Molnar succinctly summarize the options:

> "This issue of multiple truths can be addressed either by reporting all counterfactual explanations or by having a criterion to evaluate counterfactuals and select the best one."[57]

Both routes have their drawbacks. The first option – reporting all possible counterfactual explanations – might lead to information overkill that smothers the idea of crisp guidance on how to change one's actions to obtain a more desirable decision from the ADM system. Say that a student submits an essay that is graded by a grading ADM, a so-called automated essay scoring system[58], using a complex ML-generated model. The student gets a C and wants to know what he/she should change in order to get a higher grade. What the student is looking for in a counterfactual explanation is a concise pointer with regard to how to change his/her grading classification: "Use 30% more references to academic sources" or "Make fewer grammatical errors." Instead, if the student were to be provided with a list of 200 conflicting and complex counterfactuals, he/she would probably still be confused as to why the ADM system classified his/her essay as deserving of a C instead of a higher grade. A second

56 Dandl & Molnar (2019), p. 242. See above, footnote 45.

57 *Ibid.*

58 Crossley, S. (2020). Linguistic features in writing quality and development: An overview. *Journal of Writing Research, 11*(3); Dasgupta, T., Naskar, A., Dey, L., & Saha, R. (2018). Augmenting textual qualitative features in deep convolution recurrent neural network for automatic essay scoring. *Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications*; Uto, M., & Okano, M. (2020). Robust neural automated essay scoring using item response theory. *International Conference on Artificial Intelligence in Education*.

problem with an exhaustive list of counterfactuals is that it might disclose so much information about the model underlying the ADM that its owner or creator fears that the whole model can be deduced, which would raise traditional fears about the system being gamed and trade secrets being revealed.[59]

If an exhaustive list is unattractive, it might be better to make a selection of counterfactuals that are presented. Barocas et al.[60] argue convincingly that this makes counterfactual explanations subject to all kinds of undesirable hidden decisions and assumptions. For instance, one could remove all counterfactuals that are not actionable; attributes like gender, age and ethnicity are not easily changed. Does that mean it is better not to bother the subject of an ADM decision with these counterfactuals? Or would that simply be a form of window-dressing?

One could imagine a scenario where postal code and prior attainment of students from the same school are the most important factors in an ADM-based grading decision, but because these are static characteristics, a paternalistic counterfactual explanation would only point to some factors of lesser impact (grammatical errors and percentage of academic references). This might result in an individual struggling to adjust on the basis of this counterfactual, but having these efforts obliterated by factors that are left out of the picture. What about the reverse: being upfront about the most important factors, even if they do not allow for any action? In short, that would counteract the important rationale of counterfactual explanations: that they provide *individually actionable* transparency.

To get a *what if* explanation that tells you that the ADM decision would have been different if you had lived elsewhere or had gone to a different school only gives you the unhelpful information that you should have moved town or changed schools. What you really need is a *what if* explanation that points you to aspects that are relatively easy to alter, and that do not depend on the possession of resources that are inaccessible to most.

Counterfactual explanations might also fail to show important interactions between different variables. By changing a factor pointed out in a counterfactual, you might inadvertently change

59 Barocas, Selbst & Raghavan (2020), see above, footnote 29.

60 *Ibid.*

another factor that is also decisive. For example, your counterfactual tells the student that 30% more academic references in an essay would have yielded a B instead of a C. In writing the next essay, the student includes more academic references. However, the new essay also gets a C – this time not because of a lack of academic references, but because the proportion of irrelevant academic references has exceeded a certain crucial threshold in the grading ADM (information that did not come to the fore in the first counterfactual explanation because, given the lack of academic references, the content variable was unproblematic).

Another problem with counterfactuals is that notions like "nearest neighbor" or "closest data point that gets a different output" are dependent on how different scales are compared. Which essay is closest to the essay submitted by the student: the same essay written by a synthetic alter ego based at a different school, a very similar essay with a dozen grammatical errors removed or the same essay with ten additional references? When building counterfactual models, differences that are incommensurable have to be quantified. Such operationalization decisions preceding the mathematics have a crucial impact on which counterfactual will be presented.

Lastly, Barocas et al.[61] also point to the fact that the outcomes of ADM systems in real life will not always be simple binaries (exam passed or failed). In the example above, the student wants to know how to get a grade higher than a C. In practice, this would mean that a list of counterfactuals has to be provided for each grade that is higher – B, A and A* – making the already long and complex list of counterfactuals even longer and more complex.

In conclusion, it is important to point out something I have also argued elsewhere: counterfactuals, like any other ML models and outputs, are the results of constructivist processes.[62] This is not inherently negative. Constructions can be good or bad: I am not against traveling in a plane because it was constructed with a basis in a wide range of design decisions. It is only when *bad* design deci-

61 *Ibid.*

62 De Vries, K. (2013). Privacy, due process and the computational turn. A parable and a first analysis. In M. Hildebrandt & K. De Vries (Eds.), *Privacy, Due Process and the Computational Turn. The Philosophy of Law Meets the Philosophy of Technology* (pp. 9–38). Routledge; De Vries, K. (2021). Transparent Dreams (Are Made of This): Counterfactuals as Transparency Tools in ADM. *Critical Analysis of Law, 8*(1).

sions have been made, that make the plane prone to crashing, that I object. Constructivism is good, as long as it is not hidden from sight. However, when counterfactual explanations are presented as if they could not have been different – as natural truths – things do become problematic.

Counterfactual explanations are promising tools. However, for a variety of reasons, including the Rashomon effect, counterfactual explanations are no panacea against all algorithmic suffering. The individual might ultimately be confronted with a fundamental opacity because the algorithm moves in mysterious ways. While counterfactuals can be empowering for affected individuals, they are prosthetic constructions that will often be built on top of algorithmic-bureaucratic decision-making systems that are not inherently engaged with individual concerns.

6 How to think in terms of algodicies and counterfactuals in (legal) practice

The reader of this volume, *The Nordic Yearbook of Law and Informatics*, might be bewildered: where is the *legal* aspect of this contribution? What do algodicies and counterfactuals have to do with *law*? No law requires a high-level *algodicy* or an individualized *counterfactual* to justify an ADM-based decision in the public sector. This is not to say that there are no legal requirements with regard to ADM-based decisions. As mentioned above, in section III, EU data protection law (the GDPR) gives individuals subjected to automated individual decision-making with legal or similarly significant effects (Art. 22) the right to obtain "meaningful information about the logic involved, as well as the significance and the envisaged consequences" (Art. 15(1–h)). This right, which some have coined the "right to explanation" in relation to ADM[63] does not specify what kind of meaningful information is required. National administrative law in most member states requires that decisions issued by public authorities, which might be human-made or ADM-generated, have

63 Selbst, A. D., & Powles, J. (2017). Meaningful information and the right to explanation. *International Data Privacy Law 7*(4), pp. 233–242.

some kind of justification. For example, 32 § of the Swedish Administrative Procedure Act[64] reads:

> "A decision that can be expected to affect someone's situation in a not insignificant way shall contain a clarifying statement of reasons if this not obviously unnecessary. The statement of reasons shall contain information about what provisions have been applied and what circumstances have been decisive for the position taken by the authority."

Art. 3:46 of the Dutch General Administrative Law Act[65] reads:

> "A decision must be based on sound reasons (…)."

Lastly, the recently proposed AI Act[66] states in Art. 13:

> "High-risk AI systems shall be designed and developed in such a way to ensure that their operation is sufficiently transparent to enable users to interpret the system's output and use it appropriately. An appropriate type and degree of transparency shall be ensured, (…)."

Legal requirements with regard to the justification of decisions are often vague: "meaningful information," "a clarifying statement of reasons," "sound reasons" or "an appropriate type and degree of transparency" could be almost anything.

The introduction of the neologism *algodicy* aims to give a practical reminder to public authorities planning to use ADM, that algorithmic-bureaucratic practices tend to operate at a level that prevents them from engaging with the justification of decisions at an individual level. This tendency can be further strengthened when the ADM is based on a complex, ML-generated algorithm. Before using ADM in the public sector, a useful checklist of questions to raise could be:

64 Förvaltningslag (2017:900).

65 Wet van 4 juni 1992, houdende algemene regels van bestuursrecht (Algemene wet bestuursrecht).

66 Proposal for a Regulation of the European Parliament and of the Council laying down harmonized rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative Acts, Brussels, 21.4.2021, COM(2021) 206 final (2021).

– What is the purpose of the ADM system? Does it have a population management purpose, i.e., steering the population as a flock[67] for which the *overall* value and well-being should be maximized, or does it aim to address individual concerns? Does the realization of this purpose interfere with individual rights? Is collateral damage to individual rights to be expected, and is such damage legitimized?
– Is the ADM system based in statistical generalizations? Are characteristics from some individuals generalized to other individuals? Is previous behavior from the affected individual or other individuals projected on the current situation? Does the ADM system reinforce a historical status quo?
– Can the affected individual get individualized and actionable information with regard to the ADM decision?
– Is there a clear and accessible route to appeal the ADM-based classification or decision?
– To what extent is the overall justification of the ADM system a (too) high-level algodicy?

The discussion of *counterfactuals* in this text serves as a concrete elaboration of what individualized and actionable information could look like. It also serves as an example of one possible tool to satisfy relatively vague legal requirements with regard to meaningful information, transparency, the reasons underlying administrative decisions, etc. I argue that counterfactuals are promising tools in contexts where ADM decisions and their potential negative consequences need to be justified in relation to individuals. However, for a variety of reasons, including the Rashomon effect, counterfactual explanations are no panacea against all algorithmic suffering, and the individual might ultimately be confronted with a fundamental opacity because the algorithm moves in mysterious, or at least in not fully transparent, ways. While counterfactuals can be empowering for affected individuals, they are prosthetic constructions that will often be built on top of algorithmic-bureaucratic decision-making systems that are not inherently engaged with individual concerns.

67 Foucault, M. (2000). "Omnes et Singulatim": Towards a Critique of Political Reason. In J. D. Faubion (Ed.), *Power. Essential works of Foucault, 1954-1984* (pp. 298–325). Allen Lane.

7 Afterthoughts: algodicies and counterfactuals beyond the case of individual suffering resulting from a grading ADM.

In this text, I have elaborated on the notions of algodicies and counterfactuals in relation to the use of ADM systems to make grading decisions and the individual suffering that can result therefrom. The relevance of the notions of algodicies and counterfactuals is not limited to the area of grading; they can apply to any other area in which ADM is used by public authorities. Three questions deserve some additional attention in this concluding part of the text.

The first afterthought concerns the question if the notions of *algodicy* and *counterfactual* could also have relevance in the context of *human* bureaucratic decision-making. After all, as I argued above, bureaucratic decision-making machineries have a tendency to overlook individual concerns even without any ADM being involved. Thus, algodicies and counterfactuals might be *partially* relevant in human bureaucratic decision-making, but there are some salient differences. I can clarify by taking myself and my human grading decisions as an example. As a university teacher, I make grading decisions on a regular basis. While I have no difficulty distinguishing the extremes (the hopeless and the brilliant submissions), the grey zone in the middle is more demanding. There is no objectively just grade, and framing plays an important role. Elements that are important in the grading decision include the grading scale,[68] the grading criteria and the type of test, and the format of the required justification. After having worked in a particular academic setting and with a particular grading scale for a while, I tend to develop a quick gut feeling for a grade (an internalized, implicit and probably highly complex grading model). This feeling is then fine-tuned by assessment through formal criteria, comparisons, consultations with other teachers and challenges by students. A point of departure is that the work of students should be assessed *individually*, not to maintain a certain grade division. The goal is to make a fair assessment of the work of each individual student, even if that can result in every student in a class getting an A* or failing the exam. Wholly

68 Grading scales vary widely. For example, education institutions in the Netherlands work with a 10-point scale, in Denmark with a 7-point scale, and in Sweden with a 2- (pass/fail), 3- or 4-point scale.

individualized assessment is of course an ideal. Knowing that every student taking a particular course within the last ten years got an A or A* is the type of information that is difficult to exclude completely from a grading assessment. Yet, my own individual experience – for what it is worth – would lead me to assume that at least *some* human assessors can more easily than ADM systems focus on an individual case, diverge from a historical status quo and be less limited by a high-level managerial perspective. In terms of opacity, the human decision-making process might not be as transparent as we would like to think. Many decisions might be based on gut feelings (including unwarranted biases) that are justified only in retrospect. As the UK grading controversy showed, teacher predictions of grades are far from free of biases.[69]

In this sense, counterfactuals can be equally useful in human decision-making. When it is difficult to untangle which element makes a submission good or bad, it can be helpful to think in hypotheticals: "Would it have made a difference if element X or Y had been different?" When I face a discontented student and have to explain how he/she could have avoided failing his/her exam, I also face a Rashomon effect: there is normally a whole range of counterfactuals. If the student had added a few more proper references to the text *and* had structured it better, *or* made a more coherent argument *and* fewer grammatical errors, *or* … – and so on. However, based on domain-specific knowledge and experience of acting in an academic environment and engaging with students, human teachers are more likely to know which counterfactual example might be the most enlightening and empowering for an individual student who wants to get a better grade on the next exam.

A second afterthought[70] relates to the usefulness of counterfactuals. In this text, I have argued that the main usefulness of counterfactuals lies in their capacity to act as a form of individualized and actionable information to empower individuals affected negatively by ADM. Can counterfactuals be similarly useful in other settings, for example as tools used by supervisory bodies or in the court-

69 Wachter et al. (2018). See above, footnote 44.

70 I want to thank Chris Reed and Keri Grieman for their wonderful comments that helped me develop this aspect and for inviting me to present an earlier draft of this text at Queen Mary University of London (*How to Open the Algorithmic Black Box*, 24/02/21).

room in criminal or civil liability cases regarding damages following from ADM? Counterfactuals are probably not the first weapon of choice in situations where an *entire* ADM system is put on the stand. In such situations, it is warranted to look at the functioning of the ADM at a systemic level and in technical detail, which can include looking at the training data, the source code or the creation of a simplified model on top of more complex one in order to capture the essentials of its function (see above, in section III, for more details on such systemic transparency methods). However, next to systemic transparency, counterfactuals could play a role as an intuitive method to create synthetic samples to show how predictions change across particular cases. As Wachter et al.[71] put it:

> "Principally, counterfactuals bypass the substantial challenge of explaining the internal workings of complex machine learning systems."[72]

In situations where "easily digestible" information is needed to challenge a decision "and [alter] future behaviour for a better result,"[73] it is useful to have a set of hypothetic "closest possible worlds"[74] that would lead to different outcomes. In situations where it is necessary to shed light on the internal workings of an entire ADM system, such "lightweight form of explanation"[75] will not do: there, counterfactuals can at best be supporting evidence.

A final afterthought relates to the Russian doll problem that arises when an ADM based on a complex ML model is illuminated with counterfactuals which themselves are created with a complex ML model. Is there an infinite regression, and will we need counterfactuals to illuminate why certain counterfactuals were generated and not others? In principle, it would be entirely possible to build an infinite regression of counterfactuals. However, if the presumption underlying such infinite regression is that some ultimate explanation can emerge, the enterprise would be doomed to fail. Counterfactuals are

71 Wachter et al. (2018). See above, footnote 44.

72 *Idem*, p. 860.

73 *Ibid.*

74 *Idem*, p. 845.

75 *Idem*, p. 880.

lightweight, constructivist prosthetics for ADM systems to concretize some small, individualized aspects of systems that are otherwise too complex, large or abstract to grasp. Counterfactuals offer a patch to decrease the tension between the systemic logic of an ADM and the ADM's output in an individual case, but they can never resolve the tension completely, because the dots of a complex model can be connected equally well in multiple ways.

8 References

Abdul, A., Vermeulen, J., Wang, D., Lim, B. Y., & Kankanhalli, M. (2018). *Trends and trajectories for explainable, accountable and intelligible systems: An HCI research agenda.* Paper presented at the Proceedings of the 2018 HCI Conference on Human Factors in Computing Systems.

Adams, R. (2021). Teachers face 'almost impossible task' awarding A-level and GCSE grades. Study finds students in England from graduate households received more generously assessed grades. *The Guardian*. Retrieved from https://www.theguardian.com/education/2021/jun/08/teachers-face-almost-impossible-task-awarding-a-level-and-gcse-grades.

Amoore, L. (2020). Why 'Ditch the algorithm' is the future of political protest. *The Guardian*. Retrieved from https://www.theguardian.com/commentisfree/2020/aug/19/ditch-the-algorithm-generation-students-a-levels-politics.

Barocas, S., Selbst, A. D., & Raghavan, M. (2020). *The hidden assumptions behind counterfactual explanations and principal reasons.* Paper presented at the Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency.

Bateson, G. (1972). *Steps to An Ecology of Mind: Collected Essays in Anthropology, Psychiatry, Evolution, and Epistemology*. New Jersey: Jason Aronson Inc.

Blauw, S. (2020). Algorithms are biased, but so are people. We need to decide which bias we prefer. *The Correspondent*. Retrieved from https://thecorrespondent.com/288/algorithms-are-biased-but-so-are-people-we-need-to-decide-which-bias-we-prefer/3809118297 6-664bcebe.

The Book of Job (New international version translation). Retrieved from https://www.biblegateway.com/passage/?search=Job%201&version=NIV.

Breiman, L. (2001). Statistical modeling: the two cultures. *Statistical Science, 16*(3), 199–231.

Clarke, L. (2020). Ofqual advisor: Prioritising grade inflation was a political decision. *New Statesman Tech*. Retrieved from https://www.newstatesman.com/spotlight/2020/08/ofqual-advisor-prioritising-grade-inflation-was-a-political-decision.

Crossley, S. (2020). Linguistic features in writing quality and development: An overview. *Journal of Writing Research, 11*(3).

Dandl, S., & Molnar, C. (2019). Counterfactual explanations. In C. Molnar (Ed.), *Interpretable machine learning. A Guide for Making Black Box Models Explainable*: https://christophm.github.io/interpretable-ml-book/.

Dasgupta, T., Naskar, A., Dey, L., & Saha, R. (2018). *Augmenting textual qualitative features in deep convolution recurrent neural network for automatic essay scoring.* Paper presented at the Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications.

De Vries, K. (2013). Privacy, due process and the computational turn. A parable and a first analysis. In M. Hildebrandt & K. De Vries (Eds.), *Privacy, Due Process and the Computational Turn. The Philosophy of Law Meets the Philosophy of Technology* (pp. 9–38). London: Routledge.

De Vries, K. (2020). AI policy in the Netherlands: More focus on practice than principles when it comes to trustworthiness. In S. Larsson, C. Ingram Bogusz, & J. Andersson Schwarz (Eds.), *Human-Centred AI in the EU: Trustworthiness as a strategic priority in the European Member States* (pp. 132–157). Brussels: European Liberal Forum asbl.

De Vries, K. (2021). Transparent dreams (are made of this): Counterfactuals as transparency tools in ADM. *Critical Analysis of Law, 8*(1).

Esposito, E. (2013). Digital prophecies and web intelligence. In M. Hildebrandt & K. De Vries (Eds.), *Privacy, Due Process and the Computational Turn. The Philosophy of Law Meets the Philosophy of Technology* (pp. 121–142). London: Routledge.

Esposito, E. (2018). *Future and uncertainty in the digital society*. Paper presented at the Federal Agency for Civic Education (bpb), Humboldt Institute for Internet and Society (HIIG), Berlin. https://www.bpb.de/mediathek/266822/elena-esposito-future-and-uncertainty-in-the-digital-society.

European Parliamentary Research Service. (2019). *A governance framework for algorithmic accountability and transparency.* Brussels. Retrieved from: https://www.europarl.europa.eu/RegData/etudes/STUD/2019/624262/EPRS_STU(2019)624262_EN.pdf.

Foucault, M. (2000) "Omnes et singulatim": Towards a critique of political reason. In (James D. Faubion ed.) *Power. Essential works of Foucault, 1954–1984*. London: Allen Lane.

Foucault, M., Senellart, M., & Davidson, A. I. (2007). *Security, Territory, Population: Lectures at the Collège de France, 1977–78*. Basingstoke: Palgrave Macmillan.

Frederik, J. (2021). *Zo Hadden We Het Niet Bedoeld. De Tragedie Achter de Toeslagenaffaire*. De Correspondent.

Freud, S. (1905). *Wit and Its Relation to the Unconscious*. London: Kegan Paul.

Freud, S. (1920). *A General Introduction to Psychoanalysis* (G. S. Hall, Trans.). New York: Boni and Liveright.

GDPR (General Data Protection Regulation) Regulation 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC.

Hern, A. (2020). Ofqual's A-level algorithm: why did it fail to make the grade? *The Guardian*. Retrieved from https://www.theguardian.com/education/2020/aug/21/ofqual-exams-algorithm-why-did-it-fail-make-grade-a-levels.

Hernandez, J. (2018). Making AI interpretable with generative adversarial networks. *Medium*. Retrieved from https://medium.com/square-corner-blog/making-ai-interpretable-with-generative-adversarial-networks-766abc953edf.

Hussain, D. (2020). 'Great relief' from school chiefs over government A-level grade u-turn after ministers heeded calls from Eton College headmaster to dump the 'unfair' algorithm. *Daily Mail*. Retrieved from https://www.dailymail.co.uk/news/

article-86354439/Eton-College-headmaster-leads-calls-government-scrap-unfair-level-algorithm.html.

Kahn, R. (2018). Omens and algorithms: A response to Elena Esposito. *Digital Society Blog. Making sense of our connected world*. Retrieved from https://www.hiig.de/en/omens-algorithms-response-elena-espositos-future-uncertainty-digital-society/.

Lamont, T. (2021). The student and the algorithm: how the exam results fiasco threatened one pupil's future. *The Guardian*. Retrieved from https://www.theguardian.com/education/2021/feb/18/the-student-and-the-algorithm-how-the-exam-results-fiasco-threatened-one-pupils-future.

Lee, M.W, Stringer, N. & Zanini, N (2020) *Student-level equalities analyses for GCSE and A level* (Ofqual report 20/6713). Retrieved from: https://www.gov.uk/government/publications/student-level-equalities-analyses-for-gcse-and-a-level.

Leibniz, G. W. (2007). *Theodicy. Essays on the Goodness of God, the Freedom of Man and the Origin of Evil*: BiblioBazaar.

Nassar, M., Salah, K., ur Rehman, M. H., & Svetinovic, D. (2020). Blockchain for explainable and trustworthy artificial intelligence. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 10*(1), e1340.

Ofqual. (2020). *Awarding GCSE, AS, A level, advanced extension awards and extended project qualifcations in summer 2020: interim report*. Retrieved from https://www.gov.uk/government/publications/awarding-gcse-as-a-levels-in-summer-2020-interim-report.

Pasquale, F. (2015). *The Black Box Society: The Secret Algorithms That Control Money and Information*: Harvard University Press.

Proposal for a Regulation of the European Parliament and of the Council laying down harmonized rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative Acts, Brussels, 21.4.2021, COM(2021) 206 final (2021).

Raad van State (Council of State) (2018). Judgment passed on 17 January 2018, case nr. **201703951/1/A2**. Online available at: https://uitspraken.rechtspraak.nl/inziendocument?id=ECLI:NL:RVS:2018:137.

Raphals, L. (2013). *Divination and Prediction in Early China and Ancient Greece*: Cambridge University Press.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. *arXiv preprint nr. 1602.04938*. Retrieved from http://arxiv.org/abs/1602.04938.

Rimfeld, K., Malanchini, M., Hannigan, L. J., Dale, P. S., Allen, R., Hart, S. A., & Plomin, R. (2019). Teacher assessments during compulsory education are as reliable, stable and heritable as standardized test scores. *Journal of Child Psychology and Psychiatry, 60*(12), 1278–1288.

Selbst, A. D., & Powles, J. (2017). Meaningful information and the right to explanation. *International Data Privacy Law, 7*(4), 233–242.

Seligman, M., & Maier, S. (1967). Failure to escape. *Journal of Experimental Psychology, 74*, 1–9.

Sloterdijk, P. (1988). *Critique of Cynical Reason*. Minneapolis: University of Minnesota Press.

Uto, M., & Okano, M. (2020). *Robust neural automated essay scoring using item response theory.* Paper presented at the International Conference on Artificial Intelligence in Education.

Wachter, S., Mittelstadt, B., & Russell, C. (2018). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology, 31*(2), 841–887.

Waltl, B., & Vogl, R. (2018a). Explainable artificial intelligence: The new frontier in legal informatics. *Jusletter IT, 4*, 1–10.

Waltl, B., & Vogl, R. (2018b). Increasing transparency in algorithmic- decision-making with Explainable AI. *Datenschutz und Datensicherheit – DuD, 42*(10), 613–617.

Weale, S. (2020). U-turn on exams may create new set of problems in England. *The Guardian*. Retrieved from: https://www.theguardian.com/education/2020/aug/17/u-turn-exams-may-create-new-set-problems-england.

Whittaker, F. (2020*)* A-level results 2020: Top grades up by 2.4 percentage points. *FE Week.* Retrieved from: https://feweek.co.uk/2020/08/13/a-level-results-2020-top-grades-up-by-2-4-percentage-points/.

Transparency in Automated Algorithmic Decision-Making: Perspectives from the Fields of Intellectual Property and Trade Secret Law

JOHAN AXHAMN

1 Introduction

Increased attention is being given – by both policy makers, academics, businesses and government agencies – to several technologies that fall within a general description of artificial intelligence (AI). AI systems are typically software-based, but often also embedded in hardware-software systems.[1] Although there is at present no agreed definition of AI, the term is generally held to refer to systems that demonstrate intelligent behavior by analyzing their environment and taking action, with a degree of autonomy, to achieve specific goals. "Intelligence" then refers to the machine's imitation of the cognitive functions associated with the human brain, i.e., the ability to learn and solve problems.[2]

1 *Commission Staff Working Document: Impact Assessment Accompanying the Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonized Rules on Artificial Intelligence* (hereinafter "Impact Assessment"), European Commission, SWD(2021) 84 final (2021).

2 Celine Castets-Renard, *The Intersection Between AI and IP: Conflict or Complementarity?*, 51 International Review of Intellectual Property and Competition Law 141–143 (2020).

Areas where AI is used to an increasing extent include healthcare, farming, education, infrastructure management, energy, transport and logistics, public services, security, and climate change mitigation.[3] There is a belief that AI can help solve complex problems – that are beyond human capacity – for the public good and provide key competitive advantages to companies.

However, the same elements and techniques that power the benefits of AI can also bring about new risks, challenges, or even negative consequences for individuals and society. For example, AI used to replace or support human decision-making or for other activities, such as surveillance, may infringe upon individuals' rights, including fundamental rights.[4]

One way to mitigate the risks and negative consequences related to the use of AI is to set up ethical guidelines or even introduce legal obligations on the development and use of AI. The argument is that more "accountable" or "responsible" AI will foster trust and thereby adoption of and investments in the technology.[5] Ethical guidelines have been proposed by various actors[6], and legal obligations based on a risk-based assessment have recently been proposed by the European Commission.[7]

3 *Commission Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts*, European Commission, COM 206 final (2021) (hereinafter "AI Proposal").

4 Impact Assessment, *supra* note 1.

5 Heike Felzmann et al., *Transparency You Can Trust: Transparency Requirements for Artificial Intelligence Between Legal Norms and Contextual Concerns*, Big Data & Society (2019); Impact Assessment, *supra* note 1; *Presidency Conclusions: The Charter of Fundamental Rights in the Context of Artificial Intelligence and Digital Change*, Council of the European Union 11481/20 (2020); AI Proposal, *supra* note 3; Andrew Burt, *The AI Transparency Paradox*, Harvard Business Review (2019); Michael Linegang et al., *Human-Automation Collaboration in Dynamic Mission Planning: A Challenge Requiring an Ecological Approach*, Proceedings of the Human Factors and Ergonomics Society 50th Annual Meeting 2482–2486 (2006). Trust and its links to transparency, and its required conditions, have been studied in many social-scientific disciplines, including law, over a long period of time: see Stefan Larsson & Fredrik Heintz, *Transparency in Artificial Intelligence*, 9 Internet Policy Review 1–16 (2020).

6 See, e.g., Independent High-Level Expert Group on Artificial Intelligence, *Ethics Guidelines for Trustworthy AI*, European Commission (2019).

7 AI Proposal, *supra* note 3.

At the same time, it is commonly recognized that the concepts of accountability and responsibility are often intertwined with the concept of "transparency," i.e., that the technology or its use is transparent for individuals or authorities.[8] Demands for transparency of AI systems are related to the fact that the technology, by its nature, runs largely independently of human control, i.e., autonomously.[9] AI is therefore sometimes referred to as opaque[10] or a "black box."[11] A report by the Organisation for Economic Co-operation and Development explains this "black box" effect with the example of so-called neural networks as follows:

> Neural networks iterate on the data they are trained on. They find complex, multi-variable probabilistic correlations that become part of the model that they build. However, they do not indicate how data would interrelate. The data are far too complex for the human mind to understand.[12]

However, there is no common understanding of what "transparency" entails; it can refer to several different things (see below, section 3).

One aspect of transparency in AI that has gained increased attention is its relationship to the protection of intellectual property rights (IPRs) and trade secrets. In essence, the question is whether increased demand for mandatory (statutory) rules on transparency might threaten or even come into conflict with the protection of IPRs and trade secrets. The purpose of this contribution is to study this question.

The contribution is structured as follows. Section 2 is a description and analysis of how different IPRs can protect AI. Section 3 describes the possible intersection – conflict or complementarity –

8 Impact Assessment, *supra* note 1; and Frank Pasquale, Black Box Society: The Secret Algorithms That Control Money and Information (2016).

9 Martijn Van Otterlo, *A Machine Learning View on Profiling* in Privacy, Due Process and the Computational Turn: Philosophers of Law Meet Philosophers of Technology 41–64 (Mireille Hildebrandt and Katja de Vries eds., 2013).

10 Pasquale, *supra* note 8.

11 Impact Assessment, *supra* note 1.

12 *Artificial Intelligence in Society*, Organisation for Economic Cooperation and Development (2019).

between norms on increased transparency in AI and IPRs. Section 4 concludes the study.

2 Intellectual property rights and trade secret protection of AI

2.1 General

It is natural that those impacted (if not society as a whole) will want to know how an AI system reached a specific decision. However, to gain a competitive advantage from the system's commercial value or base, a company might wish to keep its AI or information related thereto secret or otherwise control its use.[13] IPRs and trade secret protection might provide such control. The following sections are a description and analysis of how AI could be protected by the IPRs of copyright and patents, or trade secret protection.

2.2 Copyright

Copyright protects literary and artistic works. There are at present several directives in EU law that have harmonized the protection of works in the Member States. Directive 2001/29 on copyright in the information society[14] regulates the protection of works in general, and there are also several more specific (*lex specialis*) directives that protect specific categories of works or types of uses. One example is Directive 2009/24 on computer programs.[15]

Depending on its form of expression, an AI could possibly fall within the protection of works set out in Directive 2001/29 (on works in general) or Directive 2009/24 (on computer programs).

Article 1 of the Directive on computer programs states that computer programs are protected by copyright, but also holds that "ideas, procedures, methods of operation or mathematical concepts"

13 Smitha Milli et al., *Model Reconstruction from Model Explanations*, Proceedings of the Conference on Fairness, Accountability, and Transparency (2019).

14 *Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the harmonisation of certain aspects of copyright and related rights in the information society*, OJ L 167 (2001).

15 *Directive 2009/24/EC of the European Parliament and of the Council of 23 April 2009 on the legal protection of computer programs* (Codified version), OJ L 111 (2009).

are not protected. Recital 7 to the Directive explains that computer programs include "programs in any form, including those which are incorporated into hardware," whereas Recital 11 provides that only a computer program's expression is protected. Ideas and principles of programs and their interfaces are not protected and neither is the "logic, algorithms and programming languages" underpinning those ideas and principles. Thus, algorithms as such are clearly excluded from copyright protection as computer programs. In the cases *Bezpečnostní*[16] and *SAS Institute*[17], the Court of Justice of the European Union (CJEU) held that object and source codes attracts copyright protection, but that an algorithm constitutes a computer functionality that is not protectable by copyright as a computer program.[18]

To constitute a "literary work" (other than a computer program), an algorithm must satisfy two criteria according to Directive 2001/29. First, it must be original, by being the author's own original intellectual creation. In its case law, the CJEU has elaborated that this may be exhibited via reflections of personality, creative choices, or sequences and combinations wherein authors express themselves in an original manner.[19] Moreover, the CJEU has stressed that subject-matter is not original "where the realization of a subject-matter has been dictated by technical considerations, rules or other constraints which have left no room for creative freedom."[20]

Whether or not these two requirements are met depends on the type of algorithm. Unsupervised learning algorithms and models are unlikely to be regarded as an author's own intellectual creation because the developer plays a limited role in their functioning. These algorithms run and learn without human supervision. A work cannot be an author's own intellectual creation if the author does not actually create it. An unsupervised learning algorithm therefore falls

16 Case C-393/09, *Bezpečnostní softwarová asociace*, ECLI:EU:C:2010:816.

17 Case C-406/10, *SAS Institute*, ECLI:EU:C:2012:259.

18 Cf. Katarina Foss-Solbrekk, *Three routes to protecting AI systems and their algorithms under IP law: The good, the bad and the ugly*, Journal of Intellectual Property Law & Practice, 2021, Vol. 16, No. 3.

19 See case C-5/08, *Infopaq International*, ECLI:EU:C:2009:465, and case C-145/10, *Painer*, ECLI:EU:C:2011:798.

20 Case C-833/18, *Brompton Bicycle*, ECLI:EU:C:2020:461.

short of amounting to a "work." However, supervised learning algorithms may satisfy the two aforementioned requirements.[21]

However, from a transparency perspective, there is a downside to protecting AIs (as algorithms) on the basis of Directive 2001/29 (as a work in general), rather than based on Directive 2009/24 (on computer programs). Directive 2009/24, but not Directive 2001/24, includes exceptions to the copyright protection that are relevant from a transparency perspective. The Directive includes mandatory provisions with the effect that parties subject to computer program licenses may study, observe, or test the licensed programs, including the source code, without infringing copyright.[22]

In summary, AIs (as algorithms) could be protected by copyright if they are the result of supervised learning. Such algorithms may be protected as works in general on the basis of Directive 2001/29. However – unlike Directive 2009/24 on computer programs – this Directive does not include provisions which give the legitimate user of the work (the algorithm) a right to study, observe, or test it.

2.3 Patents

Patents give inventors exclusive rights that preclude others from exploiting their inventions, and are filed as either product, method, or use claims. To gain protection, an invention must satisfy three main criteria. First, it must be novel, which requires that the invention is not part of the state of the art, i.e., that it is unavailable globally prior to a patent application filing. Second, it must have an inventive step, meaning that it is non-obvious to a person skilled in the art. Third, it must have a technical character, demonstrated by either creating a technical effect which serves a specific technical purpose or through being adapted to a specific technical implementation.

Computational models and algorithms enabling AI and machine learning are generally non-patentable, but if the patent claim consist of a method "involving the use of technical means" – such as a

21 Cf. Foss-Solbrekk, *supra* note 18. On AI and copyright in general, see, e.g., Johan Axhamn, *Copyright and Artificial Intelligence – with a focus on the area of music*, Festskrift til Jørgen Blomqvist 33–86 (Rosenmeier et al. eds., 2021).

22 See Directive 2009/24, *supra* note 15, Article 5(3) and Recitals 14 and 16. Articles 5 and 6 of the directive have recently been interpreted by the CJEU in case C-13/20, Top System, ECLI:EU:C:2021:811.

computer – a technical character is conferred on the subject-matter as a whole, enabling patent eligibility. Such patentable inventions are sometimes referred to as Computer-Implemented Inventions (CIIs).[23]

CIIs are patentable, provided that their method claims contain computer-executable steps or that they perform a certain functionality when deployed by a processor on a computer-readable medium hosting a computer program. Nevertheless, merely employing an algorithm in a computer to complete tasks is not technical enough. Even if the algorithm fulfils the technical consideration criterion, it often lacks an inventive step and novelty. Still, it is possible – albeit difficult – to obtain a patent for an algorithm, provided that it solves a technical problem in a novel manner and with concrete effects.[24]

Algorithms protected by patent law may facilitate some degree of transparency, as patent claims are publicly available.

2.4 Trade secrets

On 8 June 2016, following a proposal from the European Commission, the European Parliament and the Council adopted a directive with the aim to standardize the national laws in EU Member States against the unlawful acquisition, disclosure, and use of trade secrets. The Directive harmonized the definition of trade secrets in accordance with existing internationally binding standards in Article 39 of the TRIPS Agreement.[25]

According to the EU Trade Secrets Directive 2016/943[26], natural and legal persons may protect "undisclosed information" from being shared, acquired, or used without their consent, on three conditions. First, that the information is secret, meaning that it remains "generally" unknown to people commonly working with such information.

23 See, e.g., the EPO Guidelines for Examination related to Computer-Implemented Inventions. The Guidelines are available online via this link: https://www.epo.org/law-practice/legal-texts/html/guidelines/e/j.htm.

24 Cf. Foss-Solbrekk, *supra* note 18.

25 Agreement on Trade-Related Aspects of Intellectual Property Rights. The TRIPS Agreement is Annex 1C of the Marrakesh Agreement Establishing the World Trade Organization, signed in Marrakesh, Morocco, on 15 April 1994.

26 *Directive (EU) 2016/943 of the European Parliament and of the Council of 8 June 2016 on the protection of undisclosed know-how and business information (trade secrets) against their unlawful acquisition, use and disclosure*, OJ L 157 (2016).

Second, that its commercial value stems from its secrecy. Third, that the person managing said information has taken reasonable steps to protect its secrecy.[27] Recitals 1, 2, and 14 to Directive 2016/943 explain that the concept of "trade secrets" covers know-how, as well as business and technological information, in addition to commercial data on customers.

Given the general anonymity and commercial value of algorithms, they are eligible for trade secret protection. As trade secrets encompass all types of information, the training data and other proprietary information relating to the algorithm falls within this ambit. Individuals' personal data may also be included. Although Directive 2016/943 allows Member States to adopt national provisions surpassing those in the Directive, trade secrets generally safeguard information from unfair competition and commercial use, but do not, as elucidated under Recital 16 to Directive 2016/943, "create any exclusive rights to know-how or information."[28]

Trade secrets shield only against unlawful acquisitions, uses, or disclosures. Situations where these actions become lawful are recognized in Directive 2016/943. For instance, disclosures on public interest grounds or for the performance of administrative and judicial duties are permitted.[29] However, the Directive contains no provision on compulsory licensing – meaning that trade secrets must either be circumvented or disclosed to pave the way for algorithmic transparency. This is problematic when trade secrets constitute one reason why AI systems are opaque. As trade secrets are protected for as long as the information is secret, this means that AI systems may remain opaque for an indefinite period.

2.5 Database sui generis right

Within the EU, there is a specific – *sui generis* – protection of databases set out in the Database Directive 96/9.[30] The protection of

27 *Id.*, Article 2.1.

28 On this matter, see, e.g., The Harmonization and Protection of Trade Secrets in the EU (Jens Schovsbo et al. eds., 2020).

29 See Directive (EU) 2016/943 on the protection of trade secrets, *supra* note 26, Article 1.2(b) and Recitals 20 and 21.

30 *Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases*, OJ L 77 (1996).

databases is known as the *sui generis* right – a specific property right for databases that is unrelated to other forms of protection, such as copyright. Copyright and the *sui generis* right may both apply if the conditions of protection for each right are fulfilled. The Directive's provisions apply to both analogue and digital databases.

For a database to obtain *sui generis* protection, it must – according to Article 7 in the Directive – be the result of a substantial investment (measured qualitatively or quantitatively) in either the obtaining, verification, or presentation of the database contents. The protection provides the database producer with a right to prohibit extractions and reutilizations of all or substantial parts (measured quantitatively or qualitatively) of the database contents.[31]

It is debatable if data achieve *sui generis* protection under the Database Directive. The requirement of a substantial investment behind obtaining, not creating, data has caused some scholars to state that big data and AI-related data fall outside its scope.[32] However, Article 7 also grants protection to the verification or presentation of data, provided that there has been a substantial investment, either quantitatively or qualitatively. The European Commission recently noted that there is a possibility of protecting big data and data concerning machine generations and the Internet of Things as such.[33] If so, third parties are precluded from extracting substantial portions of data from owners' datasets. Consulting the database would still be possible for lawful users,[34] but beyond this any use of the database would normally be subject to a permission (license) from the rightsholder.[35]

31 The database *sui generis* right is covered by Axhamn, Databasskydd, Stockholm University (doctoral dissertation) (2016).

32 See, e.g., Anthoula Papadopoulou, *Creativity in Crisis: Are the Creations of Artificial Intelligence Worth Protecting?*, 12 Journal of Intellectual Property, Information Technology and E-Commerce Law 408, para. 1 (2021). Cf. Foss-Solbrekk, *supra* note 18.

33 See *Inception Impact Assessment: Data Act (including the review of the Directive 96/9/EC on the legal protection of databases)*, European Commission, Ref. Ares(2021)3527151 (2021).

34 Directive 96/9, *supra* note 25, Article 8, on "rights and obligations of lawful users."

35 On this matter, see, e.g., Johan Axhamn, *Databasrättens föremål i ljuset av nya förhandsavgöranden*, Nordiskt Immateriellt Rättsskydd, Vol. 78, No. 2.

3 Transparency in AI

3.1 General

Several studies and reports have highlighted the importance of increased transparency in AI. As mentioned above, increased transparency if often highlighted as part of "ethical" AI. In fact, "transparency" is the single most common ethical guideline addressing AI at a global level.[36] However, there is as yet no consensus on what increased transparency in AI would entail and research related to transparency in AI has been described to be "in its infancy."[37] There is, for example, no agreed view on which aspect of the AI that this concerns or for whom there should be increased transparency – the general public, government agencies, or citizens or other private parties?

Some scholars have described transparency as a "multifaceted concept"[38] and a "complex construct" that evades simple definitions. It can, for example, refer to inspectability, explainability, interpretability, openness, accessibility, or visibility.[39]

Some scholars make a distinction between *prospective* and *retrospective* transparency. Prospective transparency informs users about the data processing and the working of the system beforehand (*ex ante*). It describes how the AI system reaches decisions in general.[40] Thus, prospective transparency can be seen as an accountability mechanism.[41] Retrospective transparency, on the other hand, refers to *ex post ad hoc* explanations and rationales. It reveals how and why a certain decision was reached in a specific case, describing the data

36 Anna Jobin et al., *The Global Landscape of AI Ethics Guidelines*, 1 Nature Machine Intelligence 389–399 (2019).

37 Andreas Theodorou et al., *Designing and Implementing Transparency for Real Time Inspection of Autonomous Robots*, 29 Connection Science 230–241 (2017); and Larsson & Heintz, *supra* note 5.

38 Christopher Hood and David Heald eds., Transparency: The Key to Better Governance? (2006).

39 Felzmann, *supra* note 5; John Zerilli et al., *Transparency in Algorithmic and Human Decision-Making: Is There a Double Standard?*, 32 Philosophy & Technology 661–683 (2019); and Marco Tulio Ribeiro et al., *"Why Should I Trust You?": Explaining the Predictions of Any Classifier*, in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 1135/1144 (2016). Cf. Larsson & Heintz, *supra* note 5.

40 Felzmann, *id.* Cf. Larsson & Heintz, *supra* note 5.

41 Zerilli et al., *supra* note 39.

processing step by step. Retrospective transparency encompasses the notions of inspectability and explainability. Thus, for an algorithmic decision-making system to have retrospective transparency, one should be able to inspect its "internals," decompose a decision to understand the structure and weighting systems within the system, and ultimately explain a decision.[42]

A broader understanding of transparency is "access," entailing access to the AI as such.[43] As will be discussed in the next section, it is transparency in the form of "access" that poses the greatest challenges from the perspectives of intellectual property and trade secrets.

At present, norms on transparency that are relevant for AI are found in the General Data Protection Regulation (GDPR) and in non-discrimination, consumer protection, and product safety and liability rules.[44] Very recently, proposals for specific norms of transparency in AI have been put forward in the proposal for an EU Regulation on AI. The following sections will focus on the transparency provisions in the GDPR and the proposed AI Regulation.

3.2 GDPR

3.2.1 Transparency obligations in the GDPR

A transparency principle is set out in Article 5(1)(a) of the GDPR, which states that personal data must be "processed lawfully, fairly and in a transparent manner in relation to the data subject." Transparency, as understood under this Article, includes both a prospective and a retrospective element. It is prospective, because individuals must be informed about the data processing before any processing takes place. This obligation is linked to the information duties of the GDPR: data controllers are required to provide information about themselves, the quantity and quality of processed data, the timeframe of the processing activities, and the reason for and purpose of processing.[45]

The GDPR also includes a retrospective transparency element, which refers to the possibility to trace how and why a particular

42 Felzmann, *supra* note 5.

43 *Id.*

44 Larsson & Heintz, *supra* note 5.

45 Felzmann, *supra* note 5. See also The EU General Data Protection Regulation (Kuner et al. eds., 2020).

decision was reached. There are specific provisions that are particularly debated, such as the seeming right for data subjects to obtain an explanation of the decision reached where automated processing is involved.[46] Article 22(1) of the GDPR holds that data subjects shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her. However, this main rule is subject to an exception in Article 22(2), which holds that the main rule does not apply if the decision is necessary for entering into, or performance of, a contract between the data subject and a data controller (Article 22(2)(a)), or if the decision is authorized by Union or Member State law to which the controller is subject and which also lays down suitable measures to safeguard the data subject's rights and freedoms and legitimate interests, or if the decision is based on the data subject's explicit consent (Article 22(2)(c)). It is further set out in Article 22(3) that in the cases referred to in Articles 22(2)(a) and 22(2)(c), the data controller shall implement suitable measures to safeguard the data subject's rights, freedoms, and legitimate interests, at least the right to obtain human intervention on the part of the controller, to express his or her point of view, and to contest the decision.

The provisions on automated individual decision-making, including profiling, in Article 22 are supplemented with interpretive guidance in Recital 71 to the Regulation, which highlights that the data subject has "the right to obtain human intervention, to express his or her point of view, to obtain an explanation of the decision reached after such assessment and to challenge the decision." This has led to lively discussions regarding whether or not a "right to explanation" exists in the GDPR.[47][48]

In addition, the question has been raised if a right to explanation can be inferred from the wording of Articles 13(2)(f) and 15(1)(h) of the GDPR. These Articles state that meaningful information about the logic involved, as well as the significance and envisaged consequences of the processing, must be provided to the data subjects, at

46 Cf. Larsson & Heintz, *supra* note 5.

47 Bryce Goodman and Seth Flaxman, *European Union Regulations on Algorithmic Decision-Making and a "Right to Explanation"*, 38 AI Magazine 50–57 (2017).

48 Felzmann, *supra* note 5. Cf. Larsson & Heintz, *supra* note 5.

least when decisions produce legal effects for them or significantly affect them.[49][50]

3.2.2 Relationship between IPRs, trade secrets, and the GDPR

The GDPR and the Trade Secrets Directive include provisions that take aim at the intersection of the protection of personal data and trade secret protection.

Article 9(4) of the Trade Secrets Directive holds that any processing of personal data in the context of legal proceedings dealing with trade secrets shall be carried out in accordance with the previous Data Protection Directive (95/46/EC). This Directive has been repealed according to Article 94(1) GDPR, and Article 94(2) GDPR states that all references to the Data Protection Directive shall be construed as references to GDPR.

The intersection between the GDPR and IPRs and trade secret protection is referred to in Recital 63 to the GDPR. This recital holds that "where possible, the controller should be able to provide remote access to a secure system which would provide the data subject with direct access to his or her personal data. That right should not adversely affect the rights or freedoms of others, including trade secrets or intellectual property and in particular the copyright protecting the software. However, the result of those considerations should not be a refusal to provide all information to the data subject."[51]

3.3 EU Regulation on AI

3.3.1 Transparency obligations in the proposed AI Regulation

On 21 April 2021, the European Commission proposed a new EU Regulation on AI. The AI Proposal sets out harmonized rules for the development, placement on the market, and use of AI systems in the Union following a proportionate risk-based approach.[52]

To address the opacity that make some AI systems incomprehensible to or too complex for natural persons, a certain degree of

49 Andrew Selbst & Julia Powles, *Meaningful Information and the Right to Explanation*, 7 International Data Privacy Law 233–242 (2017).

50 Felzmann, *supra* note 5.

51 AI Proposal, *supra* note 3.

52 *Id.*, Article 1 and Recital 14.

transparency is proposed for so-called high-risk AI systems. Users should be able to interpret the system output and use it appropriately. High-risk AI systems should, according to the AI Proposal, therefore be accompanied by relevant documentation and instructions of use and include concise and clear information, including regarding possible risks to fundamental rights and discrimination, where appropriate. In addition, so-called logging capabilities shall ensure a level of traceability of the AI system's functioning throughout its lifecycle that is appropriate to the intended purpose of the system.[53]

For some specific AI systems, only minimum transparency obligations are proposed. Transparency obligations are proposed to apply for systems that (i) interact with humans, (ii) are used to detect emotions or determine association with (social) categories based on biometric data, or (iii) generate or manipulate content (e.g. "deep fakes"). When persons interact with an AI system or their emotions or characteristics are recognised through automated means, people must be informed of that circumstance. If an AI system is used to generate or manipulate image, audio or video content that appreciably resembles authentic content, there should – also according to the proposal – be an obligation to disclose that the content is generated through automated means, subject to exceptions for legitimate purposes (law enforcement, freedom of expression). The purpose behind these provisions is that persons should be able to make informed choices or step back from a given situation.[54]

3.3.2 The relationship between IPRs, trade secrets, and the proposed AI Regulation

A section in the annex to the impact assessment that accompanies the AI Proposal discusses the intersection of rules in transparency and intellectual property rights. It is stated that economic operators often seek copyright, patent, and trade secret protection to safeguard their knowledge of AIs and prevent disclosure of information on the logic involved in decision-making processes, the data used for training the models, etc. It is also stressed that the increased trans-

53 *Id.*, Explanatory Memorandum. See further Articles 12 and 13 and Recital 69 in the proposal.

54 *Id.*, Explanatory Memorandum. See further Article 52 and Recital 70 in the proposal.

parency obligations will not disproportionately affect the right to an intellectual property, since they will be limited only to the minimum necessary information for users, including the information to be included in the public EU database. Any disclosure of information will have to be carried out in compliance with relevant legislation in the field, including Directive 2016/943 on the protection of trade secrets. Thus, as stressed in the annex, when public authorities and notified bodies need to be given access to confidential information (protected as trade secrets) or source code to examine compliance with substantial obligations, they are placed under binding confidentiality obligations.[55]

4 Discussion and conclusion

Many AI systems are described as being "opaque" or even "black boxes," due to the difficulty in understanding or explaining the relationship between a given input and a given output. This is part of the "autonomy" of AI. At the same time, or maybe as a consequence of this, "transparency" is often put forward as an important or even necessary aspect of AI. Transparency is believed to lead to increased accountability and thus trust, adoption, and investments in the technology. However, there are different views or understandings of what "transparency" entails. It is a multifaceted concept.

Requests for increased transparency in AI may come into conflict with intellectual property protection for AI, such as copyright, patents, database rights, and trade secrets. This article has focused on this potential conflict.

As indicated above, there are some potential conflicts between the protection of IPRs (including trade secrets) and requests for increased transparency in AI. AI as algorithms may be protected by copyright, and the current copyright regime within the EU does not entail a general "right" of users of the algorithm to study, observe, or test the algorithm. Such a right is present in relation to computer programs that are protected by copyright. It should be considered whether the current EU copyright framework ought to be updated to include a right relevant to algorithms (to the extent they are pro-

55 *Id.*; and Impact Assessment, *supra* note 1.

tected by the general, or horizontal, copyright directive 2001/29) similar to that currently recognized for computer programs.

An AI may be protected by patent law, but the requirements for such protection are high. The patent protection would be related only to the commercial use and exploitation of the AI; the AI as such would be available in the patent registry.

The datasets of an AI may be protected by the *sui generis* database right within the EU. However, lawful users of the database have the right to make extractions and re-utilizations of non-substantial parts of the database contents.

An area where there is considerable potential conflict between the demands for increased transparency and protection for information is in the field of trade secret protection. Trade secret protection protects information because and as long as it is kept secret. The recently proposed EU Regulation on AI includes provisions which make it possible for public authorities to – for reasons related to transparency – gain access to the AI, even if this would include access to information that is covered by trade secret protection. The proposed regulation serves to safeguard the interests of the holder of the trade secret, by placing the public authority under an obligation not to further disclose the information (protected as a trade secret) that it is given access to.

This study indicates that there are challenges related to the intersection between IPRs (including trade secrets) and transparency in AI that might require a "horizontal" solution. Vertical solutions within each specific IPR might not provide the necessary legal certainty and satisfy the requests (by users if AI and the public at large) for increased transparency of AI. The recently proposed EU Regulation on AI is one such horizontal instrument, incorporating several transparency obligations, while at the same time aiming to safeguard the protection of trade secrets and IPRs.

Part 3

Liability

Liability in the Era of Artificial Intelligence*

STANLEY GREENSTEIN

1 Introduction

This article examines the traditional legal notion of "liability" in the context of emerging digital technologies incorporating elements of artificial intelligence (AI). It is intended to provide a brief introduction to the notion of liability in its traditional legal form and illuminate the difficulties in applying said notion, considering the advances made concerning AI technologies. This is no easy task given the vast legal landscape: the notion of liability is treated differently depending on which legal tradition is being addressed (examples being the common law and civil law legal traditions); different countries have different approaches to liability (even within the Member States of the European Union); there are different standards for criminal law and civil law, liability is addressed in contract law, but may be addressed in a specific manner also within public international law; and, lastly, certain technologies have of late received much attention in relation to liability, such as autonomous vehicles. It is with this in mind that a cautious approach is taken, the main aim of this article being to illuminate the difficulties that can be associated with applying a traditional legal notion, which has been developed over many hundreds of years, to AI technologies.

In order to create some boundaries to this vast expanse of legal material, five concepts will be used to examine the notion of liabil-

* This article is based on a presentation given at the 35th Nordic Conference on Law and IT, 10-11 November, 2020, which had the general theme of law in the era of Artificial Intelligence (AI).

ity in relation to AI technologies. These are: "conceptualization," "control," "causation," "complexity" and "challenges," collectively referred to as "the 5 Cs." However, before addressing these, a brief examination of the meaning of liability will be performed.

2 What is Legal Liability?

A somewhat natural point of departure in examining any legal notion is to attain an idea of its meaning from regular language use. By looking up the meaning of "liability," a number of alternatives can be identified. For example, liability is described as "the state of being legally responsible for something: the state of being liable for something … something (such as the payment of money) for which a person or business is legally responsible … someone or something that causes problems."[1] Already here it becomes evident that the context within which the term is used has a bearing on its meaning. Delving a little deeper into the legal meaning of the notion of liability reveals that it is "the quality or state of being liable … something for which one is liable."[2] Another legal reference to the notion of liability is that it is "the fact that someone is legally responsible for something."[3] A broader description of the notion of liability sheds light also on its function:

> One of the most significant words in the field of law, liability means legal responsibility for one's acts or omissions. Failure of a person or entity to meet that responsibility leaves him/her/it open to a lawsuit for any resulting damages or a court order to perform (as in a breach of contract or violation of statute). In order to win a lawsuit the suing party (plaintiff) must prove the legal liability of the defendant if the plaintiff's allegations are shown to be true. This requires evidence of the duty to act, the failure to fulfill that duty and the connection (proximate cause) of that failure to some injury or harm to the plaintiff. Liability also applies to alleged criminal acts in which

1 Merriam-Webster legal dictionary, https://www.merriam-webster.com/dictionary/liability#legalDictionary (last accessed 2020-10-21).

2 Merriam-Webster legal dictionary, https://www.merriam-webster.com/dictionary/liability#legalDictionary (last accessed 2020-10-21).

3 Cambridge Dictionary, https://dictionary.cambridge.org/dictionary/english/liability (last accessed 2010-10-21).

> the defendant may be responsible for his/her acts which constitute a crime, thus making him/her subject to conviction and punishment.[4]

A word that often crops up in the context of liability is "responsibility," with the phrase "legal responsibility" sometimes being used interchangeably with "liability." It can be said that the word "responsibility" has a more general linguistic application and denotes a moral duty, whereas "liability" denotes a legal responsibility.[5] For example, one may be responsible for driving a car in accordance with the prescribed law, but if you drive your car in a manner that is deemed, say, negligent and injure someone as a result of this negligence, you may be liable for the ensuing damages. The above can be framed in a manner where in the context of AI, initiatives to regulate AI have resulted in documents that refer to "responsible AI." In a way, this indicates that these regulatory initiatives are not legally binding, but rather attempt to establish a moral norm with regard to the use of AI technology.

It is evident that there is an inherent correlation between responsibility and liability. In other words, it is expected that we humans act in a certain manner in a certain context (responsibility) and that we should expect to be held legally accountable if we do not (liability). Here, attention can be drawn to a Council of Europe study entitled "Responsibility and AI," which examines the use of AI within society and the notion of "responsibility" (within the human rights context).[6] In the study, it is stated that:

> It concludes that, if we are to take human rights seriously in a globally connected digital age, we cannot allow the power of our advanced digital technologies and systems, and those who develop and implement them, to be accrued and exercised without responsibility.[7]

4 LAW.COM https://dictionary.law.com/Default.aspx?selected=1151&bold (accessed 2020-10-21).

5 Translegal, https://www.translegal.com.cn/uncategorized/responsibility-vs-liability (accessed 2020-10-21).

6 Council of Europe, Responsibility and AI, DGI (2019)05, available at https://rm.coe.int/responsability-and-ai-en/168097d9c5 (accessed 2020-10-21).

7 Council of Europe, Responsibility and AI, DGI (2019)05, available at https://rm.coe.int/responsability-and-ai-en/168097d9c5 (accessed 2020-10-21), p. 6.

There is also a philosophical aspect to liability and responsibility. The social functions of responsibility are referred to by Watson as the "two faces" of responsibility, which encompasses dual notions. The first is the notion of being in the world and acting as moral agents and being the authors of our own lives:

> Responsibility is important to issues about what it is to lead a life, indeed about what it is to have a life in the biographical sense, and about the quality and character of that life. These issues reflect one face of responsibility (what I will call its aretaic face).[8]

The second face of responsibility concerns holding people accountable:

> But "shoddy" need not express "censure." That implies a public forum, in which the subject is liable to formal sanction. To speak of conduct as deserving of "censure," or "remonstration," as "outrageous," "unconscionable" (and on some views, even as "wrong"), is to suggest that some *further response* to the agent is (in principle) appropriate. It is to invoke the practice of holding people morally accountable, in which (typically) the judge (or if not the judge, other members of the moral community) is entitled (in principle) to react in various ways.[9]

Watson differentiates between these two faces of responsibility by means of the following scenario:

> If someone betrays her ideals by choosing a dull but secure occupation in favor of a riskier but potentially more enriching one, or endangers something of deep importance to her life for trivial ends (by sleeping too little and drinking too much before important performances, for example), then she has acted badly—cowardly, self-indulgently, at least unwisely. But by these assessments we are not thereby *holding* her responsible, as distinct from holding her to be responsible. To do that, we would have to think that she is accountable to us or to others, whereas in many cases we suppose that such behavior is "nobody's business." Unless we think she is

8 Watson, Gary, *Agency and Answerability: Selected Essays*, 2004, available at https://oxford.universitypressscholarship.com/view/10.1093/acprof:oso/9780199272273.001.0001/acprof-9780199272273, pp. 263–264.

9 Ibid, p. 265.

> responsible to us or to others to live the best life she can—and that is a moral question—we do not think she is accountable here. If her timid or foolish behavior also harms others, and thereby violates requirements of interpersonal relations, that is a different matter.[10]

This section has investigated the notion of liability from a theoretical perspective, incorporating a philosophical perspective and also examining it in relation to the notion of responsibility. With this theoretical backdrop, the next section will examine the notion of liability from a more practical perspective.

3 Why is Liability Important in the Era of AI?

It is common knowledge that liability is a central concept within law. What is gaining more attention nowadays is the fact that notions associated with the application of traditional law, such as liability, are becoming much harder to apply in the techno-centric context. This theme as such is not new and interpreting the black letter law and legal notions in the face of evolving technologies has always been a challenge. However, what is novel is the complexity of the "tools." AI technologies have a complexity far beyond the technologies that we have encountered previously. According to the White Paper on AI, liability-related issues have been identified as one of the main risks associated with AI.[11] Not only are the risks associated with AI increasing, but they are increasingly becoming hidden or embedded in products and services – the autonomous car being one such example, where an incorrectly identified object can result in damages.[12] There are potentially many risks at play. There is the risk that a prevailing legal uncertainty will prevent the AI-related technologies from being utilized to improve society.[13] There is also the fact that the technologies themselves cause harm to people,

10 Ibid, pp. 266–267.

11 European Commission, White Paper: On Artificial Intelligence – A European approach to excellence and trust, COM(2020) 65 final, Brussels, 19.2.2020, available at https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf (accessed 2020-10-21), p. 10.

12 Ibid, p. 12.

13 Ibid, p. 12.

who have no remedy because of the difficulties in applying the law to these technologies, for example due to conceptualization issues. This damage may be far-reaching, materializing not only in financial terms but also detrimentally affecting more abstract values, such as fundamental rights. The notion of liability is therefore important for a number of reasons. First, it is an established legal principle that requires protecting if law as a mechanism for solving problems is to retain its relevance. Second, as the risks with the use of AI-related tools increase, so too will increased liability be required in order for those that suffer damages to have a path of recourse.

It can therefore be argued that not only is liability as a notion important, it is also a central mechanism that allows people to enforce the rights that they have been bestowed with via the law. However, having rights is one thing and enforcing them is another. In order to be able to enforce rights, liability must be attached to a party and an action, which is not straightforward in the AI environment.

4 Attaching Liability

An initial problem in identifying responsibility and liability is the fact that AI is not yet formally legally defined. There have been many attempts to describe what it is and provide a formal definition; however, the reality remains that there is no legal definition of AI, which it is argued is a precondition for settling legal liability issues.[14]

It is also argued that the manner in which we conceptualize technologies incorporating elements of AI will determine the extent to which current legal frameworks are applicable. Once again, the White Paper mentioned above is drawn on to illuminate this argu-

14 Here, reference can be made to the European Commission draft proposal for a regulation on AI that does include a definition of "AI system". It states that AI system means, "software that is developed with one or more of the techniques and approaches listed in Annex I and can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with" (Article 3), European Commission, *Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonized Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain European Union Acts*, Brussels, 21.4.2021 COM(2021) 206 final, available at https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206, (last accessed 2021-04-29).

ment. An initial point is that current EU safety legislation applies to products and not services, thereby in effect narrowing the scope of its application in relation to products incorporating elements of AI. In addition, it is debatable whether standalone software is covered by product safety rules.[15] Another issue pertains to the material application of product safety legislation. As described above, products incorporating elements of AI, such as machine learning, are continuously updating themselves by means of their self-learning capabilities. A crucial phrase within the context of product safety legislation is that of "placing on the market." Only risks or safety defects that were present at the time of placing on the market are addressed by this legal framework. In relation to machine learning algorithms, it is argued that this phrase will be difficult to apply in many cases.

Another issue highlighted by the White Paper is that of the long and complex supply chains within the context of AI products. Traditionally, product liability law has attached liability to the producer that places the product on the market. However, uncertainty can arise as to at what point an AI component was added to a product and what the status was of the entity placing the product on the market (it may not necessarily have been the producer).[16]

5 The "Five Cs"

Considering that the legal landscape is so wide regarding the notion of "liability" in the era of AI, the perspective here will be to examine the notion through five pre-defined concepts that are deemed important in this context. These are "conceptualization," "control," "causation," "complexity" and "challenges."

5.1 The Concept of Conceptualization

The literature regarding AI and technologies incorporating elements of AI is immense and an initial observation is that there is no consensus on what is really at the center of discussions. As will be seen below, the linguistic discourse is not only paramount when attempt-

15 Ibid, p. 14.

16 Ibid, p. 14.

ing to define exactly what it is we are discussing, but also has very concrete legal consequences in relation to the notion of liability. For example, when referring to AI, terms such as "an AI," "AI agents," "autonomous agents," "emerging digital technologies" and "robots" are used, to mention but a few. The above problem is not that unusual considering that AI as such is a rather diffuse concept, functioning as an umbrella concept for many different technologies. Complicating matters is the extent to which AI is a moving target: what was considered AI fifteen years ago is now considered commonplace. The disciplinary perspective attached to various arguments is also relevant, and being a technician, a lawyer, a philosopher or a layperson will affect the arguments being made. Even within a discipline, there may be differing perspectives, where for example, lawyers from a common law background will provide different legal analyses than lawyers from a civil law background.

While possibly lacking certainty at present, the notion of AI can perhaps be referred to as a spectrum. At the one end (left, in Figure 1), we have very simple technologies, best described as calculating machines: simple inputs give rise to simple outputs through a process shaped by regular programming code. For example, a process that has previously been performed manually by a human and which is transformed to a digital process, without any capabilities that can be considered "intelligent," would fall towards the left of the spectrum. Artificial intelligence considered "weak" would typically reside here. At the other end of the spectrum is what is commonly referred to as "artificial general intelligence" or "superintelligence," where machines have equaled and even surpassed the cognitive capabilities of humans. Consequently, they have feelings, emotions and a consciousness, just as humans do. The technology referred to as "strong" AI would fall more towards this right end of the AI scale or spectrum.

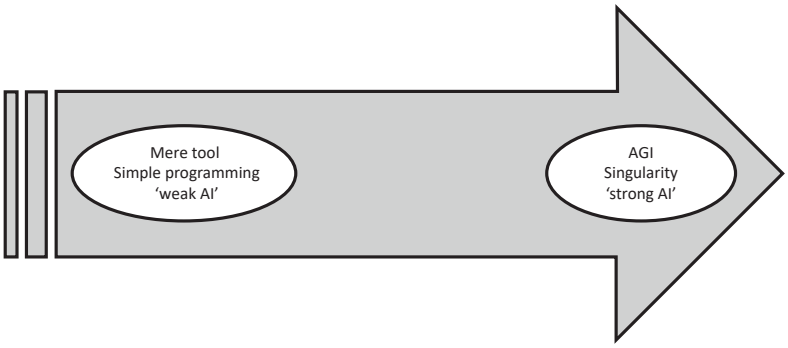


Figure 1: This simple representation depicts the spectrum of technology in relation to AI, where the technologies to the left can be described as mere tools and those to the right exhibit attributes associated with AI, such as logic, understanding, self-awareness, self-learning, emotional knowledge, planning, creativity and problem-solving capabilities. These are commonly referred to as "weak AI" and "strong AI," respectively.

Where exactly on the spectrum we should place the technologies that are available today is rather uncertain. One attribute that is important to consider is that of "autonomy." The extent to which AI agents are able to operate autonomously and in ever-changing environments will determine where on the spectrum they should be placed. One can consider autonomous robots as an example. How we conceptualize the extent to which these robots are autonomous depends, in essence, on the extent to which they are under human control, which introduces the next concept – that of control.

The main point being made here is that the extent to which AI technology is autonomous and the extent to which humans are in control at various points in the development and application of this technology will surely affect how we choose to assign liability in the traditional sense.

5.2 The Concept of Control

Bryant Walker Smith, in examining the language we use to describe various phenomena, makes the point that different actors generally use concepts differently.[17] A typical notion depicting this is the con-

17 Smith, Bryant Walker, *Lawyers and Engineers Should Speak the Same Robot Language*, in Calo, Ryan, Froomkin, A. Michael and Kerr, Ian, Robot Law, Edward Elgar Publishing, Cheltenham, 2016, at pp. 78–101.

cept of "integrity," which means something very different to lawyers than to technicians. Therefore, how we speak of technology and the phrases used to describe, for example, degrees of control, gives an indication of the autonomy of the technology – an issue that correlates the concept of control with that of conceptualization. Smith refers to conventional control theory, where systems are designed to achieve particular goals, introducing the notion of "a goal-oriented action by a subject upon an object."[18] In other words, when it comes to "regular technology" (technologies to the left in Figure 1), there is usually an external designer that imposes external goals on the technology. The problem with AI technologies is that it becomes more difficult to use certain concepts that usually denote a form of control when system boundaries are blurred (a matter discussed in the section on complexity). For example, in relation to humans, one can speak of them being "in control," "in the loop," "without control" or "out of the loop," whereas in relation to technology, one can mention that automated systems are "under control," "under human control," "under computer control," or "out of control."[19] This linguistic perspective is not foreign to the legal community, where laws and legal frameworks are required to define the extent to which control is being exerted. This is especially true where legal texts relate to the notion of liability. Therefore, the relationship between the concept of control and the notion of liability has concrete and practical legal ramifications.

An example of this is Article 22 of the General Data Protection Regulation, which regulates profiling and automated decision-making. It states that:

> The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which

18 Ibid, p. 83.

19 Ibid, p. 83. Here reference is also made to the concept of control elaborated upon in the European Commission's Independent High-Level Expert Group on Artificial Intelligence, *Ethics Guidelines for Trustworthy AI*, 8 April 2019, available at https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai (last accessed on 2021-04-29), at p. 16.

> produces legal effects concerning him or her or similarly significantly affects him or her.[20]

It can be argued that this article in essence indirectly relates to the extent of the control that a data controller has over the processing of personal data and the extent to which it can be considered "automated." Another example, provided by Smith, is of ISO 15288, which states that "Humans can be viewed as both external to a system and as system elements (i.e. operators) within a system."[21]

Thus, the concept of control is not as straightforward as with traditional technologies. As AI technologies move further to the right of the spectrum described above, it is argued that humans will exert less and less control over the technology, an issue that will require considerable debate.

5.3 The Concept of Causation

Within the legal realm, the concept of causation is a central pillar. It is common knowledge that a function of the law is to solve problems within society.[22] One way in which this is achieved is by reactively awarding some form of monetary compensation where damages have been incurred (in civil law suits) or by meting out a punishment where a crime has been committed (criminal law prosecutions). However, a central tenet of the law is that there must be a causal link between an act and the damages or crime. According to Hellner, "[c]ausation is generally an application of a general principle to a special case for a special purpose."[23] Causation has many functions, including to explain the occurrence of particular events, to predict future events, to control events, to attribute moral responsibility and legal liability and to perform certain technical applica-

20 Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), OJ L 119, 4.5.2016.

21 Supra, note 17, p. 84.

22 For a deeper discussion on the function of law, reference is made to the article by Wahlgren, Peter, that forms part of this Nordic Yearbook.

23 Hellner, Jan, *Causality and Causation in Law,* Scandinavian Studies in Law, available at https://www.scandinavianlaw.se/pdf/40-4.pdf, at p. 115.

tions of physical theories.[24] Causal chains are important in assessing liability, yet may not be straightforward. For example, event A may cause B and B may cause C, but that does not necessarily mean that C can be attributed to event A. Many different methods of solving causation issues have been incorporated into the law over the years. However, as we shall see, these may not be able to deal with the complexity of AI technologies.

Now, the link between a victim's harm and a defendant's sphere of responsibility is essential for claiming liability, where the victim has suffered damage as the result of an action. Consequently, at a pragmatic level, experiencing a harm that leads to damages is one thing, but being able to prove it is a very different matter. The victim must prove that the damage originated from the defendant's conduct or risk attributable to the defendant, and the victim will need to bring evidence of the causal connection. This is where the difficulty lies, as well as where the concept of complexity becomes relevant. Essentially, as AI technologies become more complex, it will be less evident exactly what caused damage: the sequence of events may be less evident, there may be multiple factors that are relevant in connection with the damage and who was actually in control of what part of the technology may be rather blurry.

The notion of the standard of proof can be described as the degree to which a court must be persuaded in order to hold an assertion true. This standard, however, can be very different from one country to another (even within the same legal tradition) and whether it pertains to the civil law or common law legal traditions will naturally play a role. Achieving the standard of proof required to prove a claim can therefore be extremely challenging. There are, however, indications that the situation in the European Union is changing in this respect. For example, while as of now the burden of proof is still on the victim, there are indications in EU Member States that the burden of proof can be alleviated under certain circumstances – where the exact sequence of events cannot be proven, for example, by means of prima facie evidence.[25] In some countries, the burden

24 Ibid, at p. 115.

25 European Commission, *Liability for Artificial Intelligence and Other Emerging Digital Technologies*, Expert Group on Liability and New Technologies – New Technologies Formation, 2019, available at https://data.europa.eu/doi/10.2838/25362 (last accessed on 2021-04-29).

of proof has been shifted completely, for example, for medical malpractice, see § 630h of the German Civil Code.[26]

A complicating factor in determining a causal chain of events can also be attributed to the manner in which humans cognitively conceptualize causation, as opposed to the reality of the technology. For example, while the somewhat restricted human cognitive approach views causation as a linear process, the technological processes making up AI agents are not always linear. In other words, due to the complexity of the technology and its inaccessibility to human cognitive abilities, its operation cannot be broken down into linear events and their causal effects.[27] The causes of damages may occur so closely in time that separating them would be practically impossible. This would be particularly relevant in circumstances where the multiple causes occurring so closely in time each falls under the responsibility of a different legal entity. The notion of inseparable causes is also illuminated by Karnow:

> No surgery can separate these inextricably entwined causes. No judge can isolate the "legal" cause of injury from the pervasive electronic hum in which they operate, nor separate causes from the digital universe which gives them their mutable shape and shifting sense. The result is a snarled tangle of cause and effect as impossible to sequester as the winds of the air, or the currents of the ocean. The law may realize that networks of intelligent agents are not mysterious black boxes, but rather are purposeful, artificial constructs. But that will not solve the problem of legal liability. The central doctrine of proximate cause, essential in the sorting out of multiple causes and tagging some in accordance with public policy, is useless when causes cannot be sorted out in the first place.[28]

Consequently, in attempting to establish liability, the concept of causality gains a certain importance and the above illustrates that this concept too is difficult to address in a context characterized by

26 Ibid, at p. 22, footnote 49.

27 In this regard, reference is made to the article of Chris Reed, Keri Grieman and Joseph Early in this Nordic Yearbook.

28 Karnow, Curtis E.A., *Liability for Distributed Artificial Intelligences*, Berkeley Technology Law Journal, Vol. 11:1, 1996, pp. 147–204, at p. 192.

technologies incorporating elements of AI. Another concept relevant in relation to the establishment of causation is that of complexity, which is addressed next.

5.4 The Concept of Complexity

These days, it is common knowledge that the technology of AI is becoming increasingly complex. There are no doubt multiple reasons for this, for example, developments in mathematics, the increasing access to large amounts of data and access to technologies associated with extracting knowledge using data. However, despite the technological advances, one should not lose sight of the original goals of AI, namely, the creation of machines that resemble human intelligence and have the abilities to self-replicate, to learn and to control their environment. Intelligent systems prevalent in nature were used as inspiration for the development of these machines, such as the physiological composition of the human brain, human learning processes and even the notion of evolution. In a nutshell, machines replicating human intelligence were based on the characteristics of biological entities that exhibit intelligence.

This is why the pioneers of AI were interested not only in mathematics, but also other disciplines, such as biology. This in turn gave rise to a sub-discipline within AI, namely evolutionary computing. It is here that we find different flavors of AI inspired by biology, such as neural networks and machine learning. At the core of AI is the mathematical algorithm. One type of algorithm, taking its inspiration from evolutionary computing, is the genetic algorithm.

The person credited with the development of the genetic algorithm is John H. Holland, who presented this technology in 1975 in a book entitled *Adaption in Natural and Artificial Systems*.[29] Already during the 1950s and 1960s, evolution was studied as an optimization tool and the idea was to develop a population of candidate solutions to a given problem using operators inspired by natural genetic variation and natural selection ("survival of the fittest").[30] Charles

29 Reeves, Colin, *Genetic Algorithms*, in Handbook of Metaheuristics, 2020, at p. 55, available at https://www.researchgate.net/publication/226462334_Genetic_Algorithms (last accessed 2021-04-30).

30 Ogunyale, Kuhinde, *Understanding the Genetic Algorithm*, available at https://medium.com/@kennyrich/understanding-the-genetic-algorithm-4eac04a07a59 (last accessed 2021-04-30).

Darwin's theory of natural selection formed the basic idea on which this technology is based and techniques used to get these algorithms to achieve their goals are based on principles such as "heredity" (a process that must take place whereby the children inherit some characteristics from their parents), variation (there must be diversity in the traits of the population or ways in which variation or diversity can be introduced into the population), and "selection" (some parents must pass down their genetic characteristics to the next generation and some must not).[31]

The main reason for referring to genetic algorithms as an application of AI is because it is precisely this type of technology, based on characteristics that allow for self-adaption to the environment, that is starting to make its way into more complex AI agents, especially ones that are required to operate in complex environments. Genetic algorithms, encompassing the evolutionary traits of "survival of the fittest," use various mechanisms to update themselves with minimal human intervention. They are constantly evolving of their own accord, using mechanisms such as "crossover," "mutation" and "inversion," essentially to reinvent themselves in order to find the best solution for the problem at hand, or – put differently – the goal which has been assigned to them. Notable here is the autonomy that is being conferred upon AI agents in order for them to achieve their assigned goal, something which is indicative of the shift in the approach towards AI. Here, a shift can be seen – from using "good old-fashioned" AI, which uses a symbolic approach, to an approach commonly referred to as "scruffy AI," where mechanisms of AI are based on biological and natural processes, required for adaption and learning.[32] It is within the context of autonomy that this technology gains a level of complexity that is difficult to assign traditional legal notions to. Associated herewith is the speed at which this technology is performing its operations, which is beyond human cognitive capabilities.

The above technologies are used in the context of AI because they are efficient for solving complex problems and can easily adapt to an evolving environment. It would be far too difficult and complex for

31 Ibid.

32 Brownlee, Jason, Clever Algorithms: Nature-Inspired Programming Recipes, at p. 4.

humans to program such technology manually; a correlation to the above concept of control. Thus, complexity ensues:

> We have neither pre-designed the behaviors of the robot, nor have we intervened during evolution. The robot itself and alone has developed … a set of strategies and behaviors as a result of the adaption to the environment and its own body… it is difficult to predict the robot behavior, due to the non-linearities and the feedback connections exploited for optimal navigation and obstacle avoidance.[33]

The main characteristic of this type of technology is that it is unpredictable, which is essential for operating in an unpredictable environment – autonomous vehicles being a typical illustration. Such technologies are becoming highly integrated with the environments that they are required to operate in; this is essential when the number of inputs that must be dealt with from the environment is large and they arise at a rapid pace. It can be described in terms of a system boundary becoming so porous that eventually it is difficult to distinguish the system from the environment that it is operating in. It is in this context of unpredictability that it becomes difficult to assign legal responsibility to the operators of the technology. However, "unpredictability in [the system's] operations is a feature and not a bug."[34]

The problem of complexity is eloquently addressed by Karnow, who states the following:

> Negligence and strict liability were born and raised in a Newtonian universe, the universe of billiard balls hitting billiard balls, car hitting cars; force, mass and reaction; and machinery executing one step at the time. The risk discernible in this world are the consequences of Newtonian mechanics, which is linear: A causing B causing C … With autonomous robots that are complex machines, ever more

33 Floreano, Dario, Mondada, Francesco, Cliff, Dave, Husbands, Phil, Meyer, Jean-Arcady, Wilson, Stewart, *Automatic Creation of an Autonomous Agent: Genetic Evolution of a Neural Network Driven Robot*, 1994, available at Download citation of Automatic Creation of an Autonomous Agent: Genetic Evolution of a Neural Network Driven Robot (researchgate.net) at p. 5.

34 Millar, Jason and Kerr, Ian, *Delegation, Relinquishment, and Responsibility: The Propect of Expert Robots*, in Calo, Ryan, Froomkin, A., Michael and Kerr, Ian (eds.), Robot Law, Edward Elgar Publishing, Cheltenham, 2016, at p. 107.

> complex as they interact seamless, porously, with the larger environment, linear causation gives way to complex, nonlinear interactions … the problem is not ignorance; the problem is the limits of knowledge.[35]

Having addressed the concepts of conceptualization, control, causation and complexity, the next section examines some of the practical challenges that traditional legal frameworks will face in regulating AI technologies.

5.5 Challenges

As AI technologies develop, the extent to which traditional legal notions can be applied to them is bound to decrease. This section examines the challenges for traditional law and the notions it espouses. This is done primarily from the EU perspective. The challenges posed to traditional legal notions, such as liability, originate from the concepts addressed individually above, but also from the interplay between these concepts.

A number of the initial problems with applying the notion of liability were illuminated in the abovementioned White Paper from the European Commission. The examples mentioned included the following. First, safety legislation applies to products, but not services, and there is uncertainty as to whether AI is a product or service. A second concern was whether standalone software products were covered by safety legislation. Third, the phrase "placing on the market" is difficult to apply to AI technologies as they are in a continual state of update, even once operating in their environment. A fourth concern was in relation to the long supply chains associated with AI products, something that can make the identification of a responsible party more difficult, if not impossible.[36]

It is that complexity of the technologies and the difficulty of bringing evidence where damage has been incurred, that threatens the ability of the victim of a damage to have liability assigned to the responsible party. The White Paper provides an example in the form of the Product Liability Directive in relation to autonomous cars.

35 Karnow, Curtis E.A., *The Application of Traditional Tort Theory to Embodied Machine Intelligence*, from the selected works of Curtis E. A. Karnow, available at https://works.bepress.com/curtis_karnow/9/, at p. 15.

36 Supra, at note 11.

The above directive states that a manufacturer is liable for damage caused by a defective product, however, in the case of autonomous cars, there are two main issues: 1) it may be extremely difficult to prove that there is a defect, the ensuing damage and the causal connection between the two, and 2) there may be uncertainties as to whether the directive even applies, for example, where the defect has resulted from a lack of cybersecurity robustness in the product.[37]

Another document at the EU level, also provided by the European Commission, is a report entitled *Liability for Artificial Intelligence and other Emerging Digital Technologies,* produced by the Expert Group on Liability and New Technologies – New Technologies Formation.[38] This report is essentially an overview of the extent to which legal regimes within the Member States are harmonized in the application of liability rules to emerging digital technologies. The point of departure of this report is that AI technologies may possibly cause harm, for example, bodily injuries and damages, and victims may need to seek compensation. In most Member States, this is done either by means of tort law (private law) and possibly in conjunction with insurance. Tort law is largely unharmonized, with the exception of product liability law in the form of Directive 85/374/EC.[39] The main conclusion from the expert group was that the adequacy of existing liability rules in relation to emerging technologies was questionable due to their being formulated decades ago and because they incorporate a monocausal model of harm infliction. The following characteristics of the technology were also considered to affect the ability of victims to seek compensation: its complexity, its ability to modify itself (being self-learning and continually being updated), its limited predictability and its vulnerability, for instance to cyber threats. The main conclusion was the identification of the need for an update of national liability regimes.[40]

Some of the suggestions for updating Member States' legal regimes included the following: high-risk technology should be subject to

37 Ibid, at p. 13.

38 Supra, note 25.

39 Council Directive 85/374/EEC of 25 July 1985 on the approximation of the laws, regulations and administrative provisions of the Member States concerning liability for defective products OJ L 210, 7.8.1985, p. 29–33.

40 Supra, at note 25.

strict liability on the part of those in control; where a service provider has a higher degree of control than the owner/user, this should be taken into account; manufacturers of products are to be liable for damage caused by defects in their products, even where the defects were caused by changes made to the product under the producer's control after it had been placed on the market; compulsory liability insurance is required; victims should be entitled to facilitation of proof (complex technologies); logging features are required; the destruction of data amounts to damage; and there is no need to give autonomous AI agents a separate legal personality status.[41]

The main contention from the expert group was that the more complex the digital technology becomes, the more difficult it will become to apply current liability frameworks. In coming to this conclusion, it illuminated issues such as a) the difficulty in identifying and proving causation, b) the difficulty in proving a duty of care where required, identifying that the duty of care that should have been upheld was in fact upheld or proving that it was not, c) the fact that current legal regimes are built upon the notion of human beings doing harm and monocausal harm, and d) alteration of the initial algorithm and self-learning capabilities (autonomy).

The aforementioned White Paper, in providing a partial solution to these issues, suggested a risk-based approach, where two criteria would be relevant. The first is where the AI is applied in a sector that, based on its characteristics, can be classified as high risk (for example, healthcare, transport, energy). The second is where there is a high likelihood that the risks referred to in the first characteristic will be realized (not every process within the healthcare sector involves an element of high risk). It stressed that developers and deployers of AI are already subject to European legislation on fundamental rights (for example, data protection, privacy, non-discrimination), consumer protection, and product safety and liability rules and that consumers should be able to expect the same level of safety and respect of their rights whether or not a product or system relies on AI. However, some specific features of AI (for example, opacity) can make the application and enforcement of this legislation more difficult. For this reason, there is a need to examine whether current legislation is able to address the risks of AI and can be effec-

41 Ibid. This is an abbreviated list of some of the findings of the group.

tively enforced, whether adaptations of the legislation are needed, or whether new legislation is needed.[42] A development in this regard is the publication by the European Commission of a draft of a new Regulation on AI.[43]

6 Conclusions

The main aim of this paper was merely to stress the fact that there are considerable challenges to applying traditional legal notions to technologies that incorporate elements of AI. This was done using the "5 Cs" – five concepts that operate as a prism through which this issue can be addressed. These concepts are "conceptualization," "control," "causation," "complexity" and "challenges." They were chosen due to the fact that they best illustrate the complexities of applying traditional law to emerging technologies such as AI.

Technology is advancing at an astonishing speed and this is especially true of AI. It is not only the complexity of this technology that is an issue, but also that novel technical developments that enhance the ability of this technology to operate autonomously in complex environments are continually being produced. The ideology behind the invention of AI was not only to create technology – it was to create technology with an embedded intelligence that was based on the biological foundations that make up human intelligence. How far humanity is from this goal is uncertain, to say the least. What is certain is that as the technology gets closer and closer to its goal of

42 Supra, note 11, p. 10.

43 Referenced in footnote 14. Here it can be mentioned that the draft Regulation on AI does address some of the issues in relation to liability and AI. First, the context surrounding the draft Regulation on AI is that of product safety, phrases used such as "putting on the market" and "putting in to service or use" are testament to that. Second, it also addresses many of the actors that are present in the long and complex AI supply chains, e.g., "providers," "users," "importers," "distributors," "operators" and "manufacturers," clearly setting out the obligations of each of these parties. What is interesting here is the wide definition accorded to a "distributor," namely "… any natural or legal person in the supply chain, other than the provider or the importer, that makes an AI system available on the Union market without affecting its properties." (Article 3) Then, the extent to which the draft Regulation on AI will be able to fully address the complexities associated with AI systems in relation to assigning legal liability is debatable.

creating AI, it will be necessary to adapt the traditional legal notions to this new technological reality.

Irrespective of the manner in which traditional law will in the end be used to regulate emerging technologies such as AI and without speculating with regard to the form that it will eventually take, the law as a phenomenon will be called upon to perform an important balancing act – to reap the benefits of these emerging technologies incorporating elements of AI, while at the same time protecting society from its risks.

Contractual Liability when "Things Do Not Go As Planned": A Practical Perspective

CAROLINE SUNDBERG AND JESSICA TRESSFELDT

1 Introduction

One key issue when it comes to legal risk assessment prior to deployment of new AI tools is to determine the liability of the involved parties if "things go wrong." The division of statutory liability in an AI context has been discussed in several articles, seminars, and public inquiries; however, such discussions have not yet led to any legislation, at least not in Sweden. Liability in a contractual context is more often related to the respective parties' risk appetite and insurability, and the scope of the respective parties' control. A contractual liability may also transfer the statutory risk from one party to another although the first party is always liable towards the party suffering damage as a result of the product in accordance with mandatory laws (there are, however, limitations in this regard which differ between different jurisdictions, and it is not always legally acceptable to transfer penalties aimed at punishing a specific party). Despite the extensive discussion on this topic, we have noticed that determining the appropriate contractual division of liability is something that has not yet been given a great deal of attention.

The authors of this article work in private practice and have more than a decade of experience from negotiating and interpreting IT contracts. Below, we will discuss some aspects that, based on our experience and observations, need to be resolved when drafting the IT contracts of the future, taking into account the added complexity of using AI. We are also humble to the fact that AI in itself may

change the way in which contracts are concluded and that, in the future, this process may increasingly be handled by the AI tools themselves. Such tools do not fall within the scope of this discussion, but we look forward to what the future will hold in this regard.

One viewpoint is that, in an AI context, the control of the parties may not be as easy to determine as in a more traditional IT service. This is due to the fact that there are several actors involved; the nature of AI functionality may also create a need to further analyze what the appropriate contractual division of liability is. This is something that needs to be considered in each contract going forward, at least if the market practice does not change at a pace that can keep up with the development of new products. To provide the reader with an overview of the observed contractual perspective, we will provide a brief introduction to contractual liability as generally seen in the Nordic market. Based on our experience, most of these principles also apply in a more global context, but for the sake of limiting the scope of this article, we have chosen a Nordic perspective. It should also be mentioned that AI is a highly discussed topic, meaning that new legislation and proposals are expected to come. Thus, it should be noted that this article mainly considers the legislative environment at hand during the fall of 2020.

To illustrate the current market practices as per our experience and the effects that these types of liability clauses would have in an AI context, this article will be based on four theoretical risk scenarios. Using these scenarios, we will discuss whether there is a need to amend the market practices in terms of contractual liability clauses, and we will also suggest solutions to how contractual liability may be handled.

2 Negotiation of Contractual Liability in IT Contracts – a Practical Perspective

- Below is a brief overview of IT contracting based on our practical experience in contracting for IT services in the Nordic market. Firstly, the aspect that is most often negotiated and discussed before concluding an IT contract is the clause(s) stipulating the parties' liability and, in particular, the *limitation* of liability. As the contract is often used as a tool for limiting the parties' legal

risks upon its conclusion, it is unsurprising that these clauses gain a great deal of attention from a legal perspective. Limitation of liability clauses serve the purpose of protecting (usually) both parties from potential lawsuits and exorbitant damages. Unless contractually regulated, the division of responsibility is handled in accordance with civil law principles, as applicable, depending on the jurisdiction chosen as the governing law for the contract. In our experience, it is common for the parties to limit the potential impact of said principles through contractual limitation of liability clauses. However, as there is still a lack of relevant case law in this respect, discussions on these aspects are often rather theoretical and the wording of clauses is ultimately determined based on each party's position in the market and the discretion of the party most keen on reaching an agreement.

- In this regard, suppliers often seek to limit their liability, while customers seek protection for defects and non-compliant deliveries. Thus, customers do not usually accept broad limitations of liability, but instead seek clauses that stipulate explicit indemnification undertakings. The contractual liability is also closely linked to circumstances within each respective party's control and to circumstances that the party may take out insurance against, thus resulting in a lesser risk for the party.
- Limitations of liability may be set upon each instance, as a fixed amount during the term of the contract and/or as a limitation of a specific type of damages where the liability is limited and/or excluded. Nearly all IT contracts include clauses stipulating that indirect and/or consequential damage, cost, or loss (the specific wording depends on the governing law used for the contract) is explicitly excluded from the damaging party's contractual liability, with the consequence that these types of damage are not compensated. However, the parties sometimes agree that these types of damage shall in certain situations also be covered by one party (for more information, please see, for instance, the below discussion on intellectual property rights indemnity).
- Generally, the liability is also limited to a certain monetary amount, which is usually either a certain amount or the fee paid by the customer multiplied by a given factor (generally 0.5–2 times the annual contract value, 1 being the most common). It should also be noted that not all types of liability can be handled

through contractual clauses; for example, in several jurisdictions, the liability for death or personal injury caused by a party cannot be limited. Furthermore, it should be mentioned that it is often expressly stipulated that the limitation of liability agreed shall not apply in respect of the faulting party's act of gross negligence or intent (the wording "willful misconduct" is sometimes also seen, as many IT contracts are based on US/UK contracts). Instead, the general principles of tort law apply for such acts, and at least from the perspective of Swedish law, it is not considered possible to limit liability with regard to intent and gross negligence; however, due to recent case law, this may be subject to change[1] (we will leave this topic for another article).

Another topic that is often subject to negotiations is so-called "indemnification undertakings." In this article, we use the term "indemnity" to refer to an expressed obligation to compensate for some defined loss or damage by making a monetary payment and/or to take over and manage any claims related to the undertaking to be indemnified. Indemnities may arise based on the law in the relevant jurisdiction or based on an indemnity clause. For the purposes of this article, an indemnity shall mean a clause in a contract under which one party undertakes to make a monetary payment and/or take certain actions upon the occurrence of a specified event. Indemnities may be subject to limitations, but they can also be uncapped in respect of certain types of damage.

3 AI in a Contractual Context

In an AI context, the difference between the product itself and the result of the product becomes evident. For the purposes of this article, the AI functionality may be described in a simplistic way: information (input) is inserted into an AI product that processes the input (the process), and as a result of that process, a result is provided (the output). In practice, it can be difficult to understand why a certain output has been produced (the black box problem), making it difficult to determine whether the output has been affected mainly by the process or by the input, consequently resulting in difficul-

1 NJA 2017 s. 113.

ties relating to the division of risk among the parties. Furthermore, depending on the situation, there are different parties involved in the different stages of the process, which may make this division even more difficult. Taking this into account, we need to ask the question of what would happen if a product were to produce a faulty output (for instance, if a robot were to make a mistake)? Would it even be possible to determine which party would be contractually responsible for such mistake, and would the party agree to this type of contractual liability?

There is of course no obvious answer to this question. Furthermore, the question of whether a supplier would accept any liability is, of course, purely theoretical and would most likely depend on the AI product in each respective case. However, based on our assessment, it is not likely that any supplier would accept full liability and/or indemnification undertakings in situations where the supplier cannot ensure with certainty that the input provided by the customer will not have any impact on the output, as in such situations the supplier would not be able to ensure that the output will be of certain quality and/or will not breach any applicable legislation.

4 Contractual Liability Today and Tomorrow – Four Examples

4.1 Introduction

To illustrate how AI will or may change contractual liability, we will present four example scenarios below. All examples relate to the stage when the training of the AI has been completed and the product is in production and used by end customers.

It should be noted that there are several other aspects that could be discussed within the scope of this article and each scenario. Thus, the below list is not exhaustive, and the examples presented are provided only to give the reader a contractual perspective of the topic in an AI context as a basis for further discussion.

4.2 IPR Infringements

Today, contracts where customer-specific items are developed as part of the services or products procured often include clauses on ownership of the intellectual property rights ("IPR," such as patents,

design, and copyright) to such items, as well as the usage rights, which are often heavily negotiated. Particular focus is often placed on the risk of a third party making a claim against the customer, claiming that the customer's use of the product or service is in breach of the third party's IPR (an "IPR infringement claim"). In fact, it has become more or less common practice that IT contracts include an undertaking of the supplier to compensate, and often even "indemnify," the customer in the event of an infringement claim arising from the customer's use of a resource provided by the supplier. Such clauses usually also include the right and obligation of the supplier to handle the claim and any litigation proceedings. The inclusion of such provision means that the monetary risk relating to an IPR infringement claim is transferred from the customer to the supplier.

With regard to AI and IPR, a widely discussed topic is who should be responsible for infringements of IPR. Currently, there is no certainty on how this issue will be handled by the legislator, at least not in Sweden.[2] However, regardless of whether and how the legislator decides to regulate these situations, the parties may still be able to pass on such risk contractually, unless explicitly prohibited from doing so (which does not seem likely). Hence, market practice will most likely be the deciding factor in determining which party will ultimately bear the risk for such infringements also in an AI context.

In the above context, it is necessary to distinguish between (i) infringing output caused by the process and (ii) infringing output caused by the input, and also (iii) whether the process itself would be infringing any intellectual property rights. As mentioned above, the AI tool or program consists of the process. This is perhaps best illustrated by imagining that different parties will be responsible for their respective parts of the process, resulting in difficulties when deciding which party is to blame in a statutory context. This is demonstrated in respect of contractual liability in the example below.

2 It may be noted that the European Parliament resolution of 20 October 2020 on intellectual property rights for the development of artificial intelligence technologies, 2020/2015(INI) highlights some potential issues on IPR but has not yet led to any concrete legislation. This resolution is not further discussed in this article.

Example 1
Customer A produces different types of uncomplicated texts. Customer A has identified an AI software service provided by Supplier B that would make its internal processes more efficient. The service consists of an already trained AI software that can produce texts based on input provided by the customer. When the parties negotiate their liabilities towards each other, they cannot reach an agreement on how any potential infringement claims on the output should be handled and who should be responsible for them, as none of the parties deem that they have full control over the output. What would happen if it turns out that a majority of the produced material is infringing third-party IPR because it has been trained with intellectual property protected material?

There is no obvious answer to how the described situation should be handled contractually. Under today's standard clauses, the supplier would be liable and, in many cases, the supplier would also be obliged to indemnify the customer for any third-party infringement claims. However, we consider it fairly unlikely that a supplier would agree to such clause, as the supplier does not have any control over the input data and, hence, the result of the process. Since the customer does not have any control over the output, as the supplier has trained the AI model, the customer would most likely require or at least expect the supplier to accept responsibility for its own product. Otherwise, the full risk in this respect would lie with the customer. The situation becomes ever more complex if a third party has provided the input data, especially if there is no contractual relationship between the customer and the third party providing such data.

As described above, it is not likely that any party would be willing to accept liability for circumstances outside the party's own control. One solution would be to differentiate between a third-party infringement in the process, i.e., the algorithm itself, and infringement in the output, as a supplier would likely be more willing to accept liability for the former than for the latter. Nevertheless, this does not solve the issue entirely, as the process may contribute to infringement in the output if the process has been trained with other protected material.[3] Furthermore, such division is most likely only

3 See Daniel Westman, The fourth industrial revolution and intellectual property rights (*Den fjärde industriella revolutionen – en immaterialrättslig introduktion*) in Nordiskt Immateriellt Rättsskydd (NIR) (2019 no. 1 p. 147).

theoretically possible, due to the black box problem leading to difficulties in determining why a certain output has been produced.

Another clause often negotiated under commercial contracts relates to the ownership of the IPR of the result. Naturally, such rights may have great value, and it is not uncommon that both parties wish to obtain, or at least be able to freely use and commercialize, such rights. Consequently, one approach to the above issue could be to place this risk in the contract on the party obtaining the ownership to the IPR. In fact, the party that gains the most out of the contract in respect of IPR is generally the party obtaining ownership of those rights. Taking this into account, it would be reasonable that the same party would also bear the risk for any third-party infringement. This would perhaps best be combined with a stipulation that if a party can show that the other party is responsible for the cause of the infringement, that other party will be held liable.

However, due to the involvement of multiple parties, it might be impossible to establish a standard market practice that fits all types of AI. Furthermore, it is unlikely that absolute fairness in this regard can be achieved through using a certain wording in a clause. Nevertheless, the issue should be identified and discussed among the parties before a contract is concluded. This will allow the parties to understand and assess the legal risks when using an AI service.

4.3 Biased Output

Another issue frequently highlighted in respect of AI is the risk of biased results, i.e., the risk that the result of the AI tool is discriminatory, which may lead to compensation claims from victims of such discrimination. The use of AI tools may eventually result in amended laws in respect of discrimination, but for now, these risks also need to be considered and regulated in the contract. If a party uses AI tools that have been trained by another party, it might be impossible for the first party to ensure that the output is free from any bias. Furthermore, there is no certainty as to how this risk is to be divided between the parties, and the responsibility and consequences of the biased output will likely lie with the customer using the tool. Since it is also not uncommon for contracts to include a statement that the product shall comply with "applicable law," it may be possible to place some of the liability on the supplier. However, this is something that needs to be argued and assessed

on a case-by-case basis. As of today, services are usually provided "as is" and suppliers often expressly disclaim any warranty as to the products being fit for a certain purpose (sometimes, suppliers even use disclaimers stating that they do not guarantee that the product will work at all), even if the product has a recommended use and is marketed in the relevant area of usage. In other words, the prevailing practice is that, in this regard, the customer usually bears the risk.

However, as is shown in the example below, the customer may not get to determine whether a product is fit for the customer's purposes.

Example 2
Company A is in the process of implementing a new AI recruitment tool that Supplier B is providing. The tool scans thousands of CVs within minutes and suggests candidates based on Company A's company profile as well as the requirements for the open position and previously hired personnel within Supplier B's customer base. Customer A is worried that the tool might be biased and that compensation claims might be brought should any candidate be subject to discrimination. Therefore, Customer A is seeking an indemnity in the contract for such claims and also wishes that the supplier explicitly undertakes that the AI and the output will be provided in compliance with applicable laws. Supplier B objects to this and argues that the service is used at each company's own risk and that each company needs to assess the risk and whether the tool is suitable for the intended purpose. Customer A cannot understand this approach, as it is unknown for Customer A how the tool has been trained, how it processes the information, and which characteristics determine the output.

In Sweden, the right to receive compensation for discrimination is provided in the Discrimination Act (SFS 2008:567). The compensation differs from ordinary tort law by not only compensating the victim, but also seeking to prevent discrimination. In fact, the compensation consists of both an "ordinary" compensation and a "prevention" compensation, to prevent discrimination from occurring in the first place. According to the Discrimination Act, a party who violates the prohibitions against discrimination or reprisals or who fails to fulfil its obligations shall pay compensation for the discrimination. Furthermore, the preparatory works of the Act set forth that the employer is always responsible for the discrimination, even if for example a headhunting firm has been used. If the headhunting firm

has engaged in discrimination during the recruitment process, this is to be handled contractually between the parties.[4] In the context of AI tools, the same principles would most likely apply, meaning that the employer would be responsible in case of any compensation claims. Furthermore, it is unlikely that the provider of an AI tool that is controlled by input from the customer/employer would accept any liability, especially for the "prevention" compensation, when concluding a contract.

Considering that the employer will most likely be responsible for any claims (as described above), it is probable that the contract would also include such assumptions. In fact, the only way to avoid liability would be if the customer could prove that the supplier had been grossly negligent or acted with intent when training the AI tool. Although such evidence would be difficult to produce, the customer's position would improve should several other customers also be materially affected by the biased AI product. This might be the only risk-mitigating contractual solution that the customer can expect, at least in the short term.

In summary, we do not foresee any dramatic change in terms of the standard clauses to deal with biased output, as this situation would (to some extent) be covered by current standard clauses. Therefore, it is unlikely that suppliers would be willing to extend their liability.

4.4 Product Liability

Under existing EU rules on product liability, only (i) damages resulting from death or personal injury or (ii) damage to, or destruction of, any item of property other than the defective product itself, provided that the item of property is of a type ordinarily intended for private use or consumption and was used by the injured person mainly for their own private use or consumption, are covered by product liability laws.[5] This means that if a product causes property damage to a legal person, such damage is not covered by laws on product liability. Instead, this must be handled in accordance

4 Government of Sweden, *Regeringens proposition* 2007/08:98, *Ett starkare skydd mot diskriminering*, p. 137 f.

5 Council Directive 85/374/EEC of 25 July 1985 on the approximation of the laws, regulations and administrative provisions of the Member States concerning liability for defective products, OJ L 210, 7.8.(1985), Article 9.

with tort law legislation and/or contractual liability. As shown in the example below, the damages in these cases may be substantial, and the principles currently used with regard to contractual liability may not be sufficient.

Example 3
Customer A owns a bread factory. The factory has recently implemented an automated AI baking system which includes a functionality that calculates when the bread is perfectly baked. The system not only takes account of the baking time, but also the humidity in the oven and other external factors. Due to a malfunction in the system, the AI system miscalculates the baking time, resulting in the bread factory burning down completely. The fire does not result in any personal injuries, but the financial losses are substantial. The contract with the supplier of the AI system includes a limitation of liability in accordance with market practices, limiting the damages under the contract to two times the fee paid in the preceding year.

If the contractual liability clause is applied, the customer in the example above will be compensated for only a fraction of the total damage, as the service fee was relatively low. For its part, the supplier would most likely argue that the customer must be responsible for the decision to implement software in its baking process and the supplier cannot be held liable for all the damage (including property damage) caused by the malfunction.

Notwithstanding the issue illustrated above, the situation would be handled differently if it could be shown that the malfunction was caused by gross negligence or willful misconduct by the supplier. As mentioned above, a limitation of liability is not possible in the described circumstances under current market practices in Sweden. However, to prove that the actions of the supplier have been grossly negligent would, in almost any case, be difficult, for which reason the contractual clause would most likely have very little significance if enforced.

One way to handle the risk in this regard would be to assess the "worst case scenario" regarding the occurrence of a malfunction in the AI product and to also have an undertaking in respect of at least property damage. Thereafter, it would be advisable to investigate whether the identified risks are insurable and by which party. Generally, in our experience, suppliers are usually more willing to accept liability if the liability can be insured. However, this approach can-

not be applied if none of the parties can take out insurance on the interest in question and, in any case, such coverage will most often only compensate property damage and other direct damage. Thus, the risk for costs and loss resulting from indirect damage would likely still lie with the customer.

4.5 Personal Data

As is widely known, data are essential for the use of AI, both during the training phase and in the continuous use of the AI tool. Many tools are dependent on massive amounts of data. The data may either be collected internally or purchased through a third-party vendor. Due to data protection legislation, such as the General Data Protection Regulation (the "GDPR"), *personal* data cannot be transferred and/or purchased freely. Instead, there are restrictions on how the data may be processed and with whom the data may be shared. Furthermore, several obligations, such as transparency towards the data subjects, lies with the party controlling the personal data (the data controller).

The data protection legislation does not apply to anonymous information, i.e., information which does not relate to an identified or identifiable natural person or to personal data rendered anonymously in such a manner that the data subject is not or is no longer identifiable[6]. Therefore, such anonymized data are to be preferred if data are shared between several parties for the purpose of AI tools. Even if neither party to a contract would benefit from transferring personal data, it cannot be excluded that personal data could be transferred by mistake. Under current market practices, suppliers of information seldom take responsibility for such occurrences.

6 When assessing whether a natural person is identifiable, one should consider all the means that are reasonably likely to be used. To assess whether a certain means is reasonably likely to be used to identify the natural person, all objective factors, such as the costs and the amount of time required for identification, need to be considered, taking into account the available technology and technological developments at the time of the processing. European Union, Council Regulation 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC, OJ 2016 L 119/1 [hereinafter GDPR], Recital 26.

Example 4
Customer A purchases anonymized data from Supplier B to train its AI tools. The contract with Supplier B clearly specifies that the data provided should not consist of any personal data, and in line with market practices, the contract excludes any liability for incorrect data. Furthermore, the contract stipulates that Customer A is solely responsible for ensuring that any processing of data shall be conducted in accordance with applicable legislation. Customer A is worried that the data could include personal data by accident, and under such circumstances, Customer A would be an independent controller of the data once it is transferred from Supplier B. If this were to happen, even by accident, Customer A would be subject to several obligations, such as an obligation to inform the data subjects about the processing. As the personal data would be transferred by accident and perhaps without the parties' knowledge of the transfer, there is a significant risk that Customer A could be deemed to be in breach of the GDPR and be subject to, e.g., compensation claims from data subjects.

A breach of the GDPR could result in both compensation to the data subjects and penalties. Whether the penalties may be transferred to another party contractually, considering that such penalties aim to punish the party on whom the penalties are finally imposed, has been subject to extensive discussion. A common view is that clauses transferring the risk would not be upheld by the legal order (i.e., by the courts). However, the general approach with respect to compensation claims by data subjects seems to be that transferring the liability on another party contractually is acceptable. If the breach is extensive and relates to numerous data subjects, the compensation could be substantial. As the main purpose of an AI service is to provide data, it is sensible to consider how these types of situations would be handled.

The described scenario is not regulated by the GDPR. The GDPR only stipulates that the data subjects have the right to receive compensation from the controller for the damage suffered[7]. In the described scenario, both the supplier and the customer would be independent controllers, allowing the data subjects to receive compensation from them both, if both have contributed to the breach. The situation in question relates to damage caused to a third party outside the contractual relationship, e.g. it would constitute a third-party claim from

7 GDPR, Article 82.

the supplier's perspective. Such damage is generally not covered by tort laws, and it is from our experience uncommon that a supplier would accept any liability for such claims in a contract.

As mentioned above, it is logical to place the liability on the party that has control over the damaging event. However, with respect to personal data, current market practices seem to set forth that a transferring party does not accept such liability, even though the transferring party is also the one who must ensure that the data do not include personal data. If personal data are "accidentally" transferred, the error cannot be handled later by the receiving party (the customer), because if the data have been transferred, the customer will process this information and become a data controller. Therefore, one could argue that the supplier should be held liable and accept liability for third-party claims.

Another aspect to be discussed is whether the liability should be uncapped, i.e., not be limited to a specific amount. The amount of potential damages is usually difficult to foresee when the parties conclude their contract, as it will depend on several factors, such as the number of data subjects affected.

Lastly, it should be noted that regardless of whether or not the supplier is willing to accept any liability, the customer should always consider the risk related to the purchase of data from a third party in respect of data privacy legislation.

5 Conclusions

As shown in the examples in this article, existing market practices in respect of contractual liability may not be suitable for the provision of AI systems/tools. In our opinion, it is unlikely that suppliers will unconditionally accept full or even additional contractual liability compared to current market practices (also in the case of IPR infringement situations, as has been fairly common under current market practices), which generally only cover limited amounts of direct damage. To the extent that customers will require suppliers to have extended liability, we will perhaps see a more detailed division of liability where the supplier will accept liability for defects in the process/product itself. However, as is shown above, such claims would, for the most part, be subject to material evidence issues. Nevertheless, it is difficult to foresee another way of solving the matter.

Furthermore, we will most likely see that limitations of liability in respect of certain types of damages will be subject to extensive negotiations, for example with respect to third-party claims. There may even be an increase in third-party contracts which would, of course, add to the complexity. If accepted by suppliers, this would be a material gamechanger for standard clauses. Lastly, possible amendments to indemnification clauses will also be of interest, and we are curious to see whether they will become more limited, for instance by excluding liability for IPR infringements, or whether their scope will be broadened, for instance by also including property damage caused by AI.

The legislator should also consider how the division of liability could be transferred by contractual clauses and regulate such division accordingly. Further, the legislator should consider the positions of the respective parties in the field, to ensure that smaller actors are not left or forced to accept extensive liability in relation to larger actors with a better negotiating position.

As of now, nobody knows how the contractual landscape will evolve. However, the parties' ability to freely agree on the division of liability must be taken into consideration by the legislator if and when additional changes are made to the current legislation due to the deployment of AI tools. Furthermore, it is essential that all parties involved understand their respective rights and obligations and also the risks involved in using a certain AI product. Otherwise, we can be certain that the precise intentions of the parties will not be reflected in the contracts, which will ultimately result in an increased number of disputes.

Responsibility and Accountability: AI, Governance, and the Rule of Law

RICHARD SANNERHOLM

1 Introduction

Public power is becoming automated. The question of how to regulate technology is quickly turning towards how the technology we use regulates us. Automation processes test the relevance and suitability of established concepts and frameworks for good governance. As Sheila Jasanoff writes in *The Ethics of Invention*: 'Protections are needed for our digital selves, but where should the safeguards come from, and to what extent can the wine of old principles be poured undegraded into new bottles of the digital age?'[1]

This paper examines automation of public decision-making from the perspective of the rule of law, focusing specifically on legality and accountability. The rule of law requires some unpacking, which will follow shortly, but suffice it for now to say that the rule of law is a concept for minimising arbitrary power. It is argued here that the rule of law is highly relevant to the regulation of automation. This is perhaps an unnecessary point to make since no one is (openly) against the rule of law.[2] Nevertheless, it is often suggested and sometimes asserted that old solutions are ill-suited to solve new

1 S. Jasanoff, *The Ethics of Invention: Technology and the Human Future*. Norton & Company. 2016: 18.

2 Not even countries swaying far from the rule of law are openly against the rule of law. *See* S. Walker, '*Hungary and Poland to counter critics with "rule of law institute*"'. The Guardian. 28 October 2020.

problems.[3] Moreover, when the rule of law appears in regulatory discussions, it is either in the form of general descriptions, or as a checklist of detailed individual safeguards on legality, accountability, transparency, etc. However, the rule of law is an empirical concept and a complex system beyond the binarity of checklists.

For 'automated decision-making', the definition put forward by *Algorithm Watch* is employed: 'procedures in which decisions are initially – partially or completely – delegated to another person or corporate entity, who then in turn use automatically executed decision-making models to perform an action.'[4] For artificial intelligence (AI), the *High-Level Expert Group on Artificial Intelligence* (HLEG-AI), set up by the European Commission, presents the following description: '… systems that display intelligent behaviour by analysing their environment and taking actions – with some degree of autonomy – to achieve specific goals.'[5] The automation of governance ranges from support for simple and binary tasks, such as speed cameras for issuing fines, to sophisticated technology for collecting and interpreting data more independently, including in situations where the law allows scope for assessment and evaluation, for instance in cases concerning income support or social security. Of the two, the former category of automated decision-making is the most common, but – by all estimates – it is towards the latter category of AI that governance is heading.[6]

The issue is not whether or not governance should be automated, but that the process of automation warrants a broader rule of law perspective. This involves questions that are legal, but also sociolegal

3 *See* Position Paper on Denmark, Belgium, the Czech Republic, Finland, France, Estonia, Ireland, Latvia, Luxembourg, the Netherlands, Poland, Portugal, Spain and Sweden, *AI. Innovative and trustworthy AI: Two sides of the same coin.* 2020. *See also* European Commission, *White Paper. On Artificial intelligence – A European approach to excellence and trus*t. COM(2020) 65 final. 2020; *and* UNESCO, *Preliminary Report on the first draft of the Recommendation on the Ethics of Artificial Intelligence*. CL4327. 2020.

4 Algorithm Watch, *Automating Society. Taking Stock of Automated Decision-Making in the EU*. 2019: 9. *See also* Article 29 Data Protection Working Party*, Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/689*. 2018.

5 High-Level Expert Group on AI, *A definition of AI: Main capabilities and disciplines*. European Commission. 2019.

6 R. Karlsson, *Den digitala statsförvaltningen – Rättsliga förutsättningar för automatiserade beslut, profilering och AI*. 1 Förvaltningsrättslig tidskrift 2020.

and behavioural in nature. The legal questions relate to the commonly identified risks associated with automated governance relying on algorithms and codes. Who is responsible for decisions based on faulty algorithms, i.e., where algorithms rely on biased training data, or situations where the empirical world clashes with a highly functioning model? How can decisions be challenged? What happens when algorithms, which in themselves are correct, make decisions based on incorrect data?[7] The legal risks are intertwined with socio-legal and behavioural considerations. What does automation mean to an age-old concept such as the rule of law, a concept that we tend to understand through metaphors and embodiments – for instance, a courthouse, or the saying that no one is above the law, or a public servant making a decision? This way of conceptualising law is rooted in our language and minds.[8]

The development of automation typically precedes legal discussions and legislative changes. Given the large bulk of everyday cases in the fields of taxation, social insurance, or transport tariffs, automation is necessary for effectiveness and legal certainty, to minimise the risks inherent to manual decision-making. Thus, human intelligence is not only redundant; in some instances, it is not even desired. One example of redundant and unwanted human agency can be seen in the Swedish Government Agencies Ordinance which stipulates that, as a main rule in public administration, a decision must be reported before a final decision is made. Automated decision-making naturally precludes this step.[9] The legal adaptation that has taken place through the 2018 Administrative Procedure Act, which now provides a legal basis for automated decision-making, satisfies the condition of legality. But the legal change also raises deeper questions on how we are to understand the rule of law in

7 *The Standing Committee of the Parliamentary Assembly of the Council of Europe* recently proposed that the Committee of Ministers should support the elaboration of a legally binding instrument governing AI. The European Commission, *Proposed regulation from the European Parliament and of the Council. Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Act*s 2021/0106(COD) sets forth a definition of AI and detailed rules for the development, import, distribution and use of AI systems.

8 S. Larsson, *Conceptions in the Code: How Metaphors Explain Legal Challenges in Digital Times*. Oxford Studies in Language and Law. Oxford University Press. 2017.

9 *See* Government of Sweden, *Regeringens proposition 2016/17:180, En modern och rättssäker förvaltning – ny förvaltningslag*. 2017.

its new form. What does it mean when information about who has reported and who has made a decision is no longer relevant? Surely someone, somewhere, has made a decision – for example, a human has designed the algorithm that leads to the automated decision-making. Do we know who? Does it matter? Is this the 2.0 of Aristotle's adage that 'law should rule, not men', substituting law with smart machines interpreting law?

The paper is structured in the following way. Section II will discuss the rule of law, focusing on how the concept is presented in the literature and in practice. It is suggested that a teleological view is a natural starting point, which does not preclude practical refinements. Complex social concepts must be made manageable and lists and prescriptive categories extend naturally from a process of adaptation. The two subsequent sections (III and IV) each deal with central rule of law safeguards – legality and accountability – examining situations where current regulatory approaches seldom live up to the rule of law aim of minimising arbitrary power. This is followed by a section on the robustness of rule of law systems to handle automation of governance, when automation in most countries takes place at the level of municipal or local governance. A concluding section summarises the discussion and sets out a few points on how to consider and use rule of law perspectives in future regulatory discussions.

2 Rule of law

Systems of rule of law and accountability have been developed in response to the eternal problem of how to manage the coordination of relationships between individuals. Rule of law serves as a crucial rule and norm enforcement mechanism: 'the social psychological link between individual decision-makers on the one hand and social systems on the other.'[10] Considering the importance of links between decision-makers on the one hand and social systems on the other, automation of governance extends beyond the challenge of revising laws to fit automated decision-making, meeting a minimum threshold of legality; it is also a cognitive challenge. Automation calls into

10 P. E. Tetlock, *The impact of accountability on judgement and choice: Toward a social contingency model*. 25 Advances in Experimental Social Psychology 1992.

question how we talk about and make sense of the rule of law when governance is no longer performed by humans.

Checklists and teleology

It is often said that the rule of law is nearly impossible to define. This is not true simply because it is said often. Just like democracy, good governance and other social system concepts, the rule of law can be defined. The question is how strong the explanatory power of a definition is. There is no shortage of definitions of the rule of law. These range from short and concise to lengthy and complex. What unites the definitions over time is an assumption 'so shared, so assumed that is never explicitly discussed',[11] namely that the rule of law is all about virtues internal to the state's legal system. That it is primarily, and almost exclusively, to do with the legal institutions, rules and regulations of a legal order.

Following this shared assumption there is another: that the rule of law is best constructed with inclusion of various safeguards. Lawyers and legal scholars disagree on the exact composition and nature of some of the safeguards, but they all operate with virtues internal to the legal system. Thus, Lon Fuller had a list of eight principles that, properly respected, would provide law with an inner morality, though also noting what the actual content of the law is – the formal character of legal rules, whether prospective, public, general, etc.[12] Joseph Raz has a similar list and constructs the concept of the rule of law on eight principles. His list does not differ much from Lon Fuller's, except that it is more focused on the implementation of law, the legal craftmanship, placing more emphasis on judicial review, the independence of the judiciary, and access to justice.[13] Lord Bingham also had eight principles, as does John Finnis, while Jeremy Waldron has ten and Robert Summers has eighteen rule of law principles.[14]

11 M. Krygier, *What's the Point of the Rule of Law?* 67 Buffalo Law Review 3, p. 747. 2019.

12 L. Fuller, *The Morality of Law*. University Law Publishing. 2004.

13 J. Raz, *The Authority of Law. Essays on Law and Morality.* Oxford University Press. 1979.

14 For an overview, *see* J. Waldron, *Thoughtfulness and the Rule of Law*. NYU School of Law, Public Law Research Paper No. 11–13. 2011.

It is not only academics who produce checklists on the rule of law. The Venice Commission of the Council of Europe has produced two checklists, the most recent (and lengthier one) contains five main principles and a much longer set of detailed targets and indicators. The main principles are legality, legal certainty, equality before the law, non-discrimination and access to justice.[15] The World Justice Project (WJP) has a global index on the rule of law. This involves a whole methodology and, of course, a list of principles (or categories) for measuring the concept. The WJP defines the rule of law as accountability, just laws, open government, and accessible and impartial dispute resolution.[16] With a similarly practical focus, the EU has developed a rule of law mechanism for measuring the rule of law within the Union.[17] The recent – and first – rule of law report from the European Commission is a sobering read. The rule of law is fading in many European countries. Digitalisation or automation is not one of the reasons for this, but it is likely that automation will feature more heavily in future reports from the Commission.

A contrasting view to the checklist approaches is to depart from an earlier but more fundamental starting point: an end goal – to ask what the point of the rule of law is. Why the rule of law is valued and why it has endured as a relevant concept over time has to do with its main task, namely to reduce or mitigate the arbitrary exercise of power, rather than with any of the individual institutions or principles attached to it. Law is, at its core, about the exercise of power – that is: what it does.[18]

This ends-based perspective is helpful because it links the question of how to regulate emerging problems of automated governance and AI to the issue of power, with the goal of minimising the arbitrariness that often follows power. Martin Krygier has for a long time, and convincingly, argued for a teleological perspective on the rule of law. The labyrinth of lists, checklists, and rule of law

15 European Commission for Democracy through Law (Venice Commission), *The Rule of Law Checklist*. Council of Europe. 2016.

16 World Justice Project, *Rule of Law Index 2020*.

17 Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions, 2020 *Rule of Law Report. The rule of law situation in the European Union*. COM (2020) 580 final. 2020.

18 Krygier, *supra* note 11, p. 761.

organigrams cloud rather than provide clarity on what the rule of law is good for, why it should be sought in relation to governance generally, or its relevance to emerging regulatory challenges such as automation and AI.

There is a further argument for developing Krygier's point of starting with the ends rather than the means. Suggesting to the European Parliament, the European Commission or any national legislator or public agency that when they insert the rule of law into legal texts, or uphold the rule of law in practice, they should do so from a teleological point of view has its obvious limitations, since it would be hard to know where to start and where to stop in the practical business of law-making.[19]

The main merit of the teleological view is that it focuses squarely on what the rule of law is good for. This has a value in and of itself for legislators and bureaucrats. A more important, additional value is that the rule of law, to rule it, requires a collated exercise – a system action, for want of a better term, linking legislators and decision-makers on the one hand and social systems on the other. Piecemeal legislation, where individual safeguards are inserted with little thought given to how they relate to each other and how they collectively relate to the threat of arbitrary power, has obvious limitations. Eventually, legislators and public officials will need rule of law lists of principles – but this is not the place to start until there is a clear understanding of why rule of law is wanted in the first place. Moreover, when using a checklist, legislators and bureaucrats should digest the list and assess everything on it to reach a balanced decision on if a particular governance response achieves the required goal.

Automation and the importance of rule of law

The importance of rule of law in relation to the automation of public decision-making, and AI generally, is gaining attention among regional and international organisations. At a recent Council of Europe high-level conference on AI, it was concluded that since AI positively and negatively impacts 'the exercise of human rights, the functioning of democratic societies, and the rule of law', it requires

19 *See* U. Plesner & L. Justesen, *The Double Darkness of Digitalization: Shaping Digital-ready Legislation to Reshape the Conditions for Public-sector Digitalization*, Science, Technology & Human Values, 1–28, 2021.

'timely and thoughtful policy responses and must be placed at the tip of governments political agendas'.[20] It can be difficult for the individual to fully interact with an automated or semi-automated system for decision-making and 'AI tools can support trained judges, while the content and contours of the laws and the legal systems of democratic societies must remain authoritatively governed by humans'.[21] This is a strange echo reversing Aristotle's claim that he trusted the law to rule, not men, after his encounter with tyrannical rulers. Law, according to Aristotle, stood for reason and equality before the law (*isonomia*) and man for passion, the irrational, the easily subverted and perverted ruler.[22] Now, it seems, we are concerned that man is not allowed to rule, being pushed to the margins by cold, rational, automated decision-making – but where the result might be enhanced predictability. It is not unthinkable that Aristotle would have approved.

UNESCO's *recommendation on the ethics of artificial intelligence* advances that AI is not just a technological game changer, but also an anthropological disruption. Therefore, the recommendation continues, it is necessary to adopt a more proactive thinking 'beyond the traditional legal approaches, which lag behind. The proposed Recommendation should become an ethical guiding compass and a normative bedrock to build a strong respect for the rule of law in the digital word', UNESCO concludes.[23]

An additional reason for the increased attention to the rule of law in relation to AI and automated governance is that rule of law is under threat in a growing number of countries where, until recently, it was assumed to be an integral part of governance. Thus, while the rule of law was hailed as a magic bullet for a range of global problems just a few years ago, by organisations such as the UN and the EU

20 Council of Europe, *AI: Governing the Game Changer – Impacts of artificial intelligence development on human rights, democracy, and the rule of law*. Helsinki, 26–27 February 2019.

21 *Id.*

22 Aristotle, *The Politics*. (S. Everson, ed.) Cambridge University Press. 1988.

23 UNESCO Recommendation on the Ethics of Artificial Intelligence, adopted on 24 November 2021, UNESCO's General Conference, 41st session.

and by countries as politically disparate as China and Canada, there is less consensus on the importance of the rule of law today.[24]

Moreover, where the rule of law is under threat, the main tactics used by politicians and public officials are not blunt force, throwing judges in jail and conducting nightly raids led by secret police, but more subtle. In various countries, from Guatemala to Poland and Hungary, aspiring autocrats employ a tactic of undermining the rule of law through legal means.[25] It follows the basic principle of solvents, that like dissolves like. Judges are removed from the bench using technical constitutional amendments, liability laws revised to control public services, and laws on broadcasting licenses tweaked to skew the market to the oppositions' disadvantage.

Authoritarian regimes are great adapters when it comes to keeping and reinforcing power, and using digital means is no exception.[26] Automated decision-making and AI has the potential for both uses and abuses (consider China's social credit register system).[27] In most countries, automation of governance is a far cry from social credits, but the technology lends itself to any purpose and the scalability of measures that can flow from automated governance is overwhelming. However, between repressive practices of technology, and technology as the guarantor of legal certainty, 'falls the shadow'.[28]

The whole and the sum of its parts

The UN, the EU and other international agencies sometimes wield the concept of the rule of law as a charm, without really specifying what it is. The Council of Europe, for instance, confidently asserts that there is a need 'to create a regulatory framework for AI, with

24 Krygier, *supra* note 11, p. 745.

25 *See* V-Dem Institute, *Autocratization Surges – Resistance Grows: Democracy Report 2020.* 2020. *See also* T. Ginsburg & A. Huq. *How to Lose a Constitutional Democracy*. 65 UCLA Law Review, p. 78. 2018; *and* K. L. Scheppele, *Autocratic Legalism*. 85 University of Chicago Law Review, p. 545. 2018.

26 Jasanoff, *supra* note 1, p. 18. The exploitation of digital means for keeping power, or polarizing opposition, is not only a threat in authoritarian countries. *See* C. Sunstein, *The Law of Group Polarization*. 10 Journal of Political Philosophy, pp. 175–195. 2002.

27 L. Diamond, *The Road to Digital Unfreedom*. 30 Journal of Democracy. 2019.

28 T. S. Eliot, 'The Hollow Men', *Poems 1909-1925*. Harcourt, Brace & Co. 1926. 'Between the idea, And the reality, Between the motion, And the act, Falls the shadow.'

specific principles based on the protection of human rights, democracy and rule of law'.[29]

At a quick glance this seems simple enough, but what does it mean in practice – a regulatory framework for AI with *specific principles* based on the protection of rule of law.[30] Does this include all the principles on Fuller's list, which makes sense since he was primarily devoted to the making of law? Or does it include Fuller's and Raz's, to encompass legal craftmanship too? Or is it more along the lines of the Venice Commission's checklist of five principles, with its lengthier list of detailed safeguards for each? And how would the principles of legality, legal certainty, or access to justice from the Venice Commission's checklist, for example, be made part of a regulatory framework?

Individual safeguards of the rule of law do find their way into legal texts – examples include EU General Data Protection Regulation (GDPR) and the recently proposed regulation on AI from the European Parliament and the Council.[31] After all, the checklists are based on the shared assumption that the rule of law is primarily about the internal virtues of the legal system. However, when they are inserted into regulatory frameworks this is done in a way that is separated from the teleological meaning of the rule of law.

One simple reason why there is a separation of principles and goals is that the rule of law is not upheld only by the legal system, but also by the broader social and political systems. Rule of law, since it deals with power, is also conditioned by factors such as trust, political agreements, social norms, and culture. Another reason why it is difficult to achieve any type of effect depends on the use of regulation as the medium. GDPR, for example, is not an act setting out the importance of the rule of law, but deals with data protection. It is moreover a lengthy act that must harmonise with other regulatory

29 Committee on Political Affairs and Democracy, *Need for democratic governance of artificial intelligence*. Doc. 15150. Council of Europe. 2020.

30 *See*, for example, the European Commission, *White Paper. On Artificial intelligence – A European approach to excellence and trus*t. COM(2020) 65 final. 2020, which sets forth broad principles to govern AI (and similar ethics and principles focused initiative from UNESCO).

31 European Commission, *Proposal for a regulation of the European Parliament and the Council, Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Act*s 2021/0106(COD).

frameworks, and is dependent on a multitude of agencies and actors for its implementation etc. This holds true for all manner of regulatory regimes. The chances of achieving an effect when inserting rule of law principles into regulatory frameworks, it is argued, depend on the reason why they were inserted in the first place – what was the point?

When principles flow from checklists to regulation, this may serve to enhance transparency to some extent, promote accountability in some ways, and ensure legality in many ways, but may not necessarily produce an effect that successfully minimises arbitrary power. Furthermore, the consequences of greater reliance on AI and automated decision-making will probably only reveal themselves in the long term. They may be incremental, causing numerous small harms that go undetected, but that cumulatively shift or otherwise alter the fundaments of governance taken for granted over a long period of time.

An illustrative example of the risks involved when ends and means are separated is evident in the reverse engineering that the EU is now engaged in regarding the rule of law and union values. The initial criteria for conditioning membership to the EU (then the EC) are seen in the so-called Copenhagen document and in founding EU documents. Rule of law is listed as a political prerequisite for membership, together with democracy and human rights.

The rule of law was further detailed in the membership processes under the care of the Commission, but separated from the teleological meaning of the concept. Rule of law in Slovakia focused on minority protection, in Bulgaria on organised crime, in Poland on something else etc. Breaking up the concept into principles for measuring membership progress led to a technical standard setting. Milestones were reached for each of the individual principles, but the overarching goal of the rule of law was largely unaccounted for.[32] What has triggered the EU response to rule of law threats inside the Union is not disciplinary actions against judges in Poland or how judges are selected in Hungary, but a cumulative assessment of the situation as a whole – how these legal changes, in combination with constraints on civil society, education, culture and media, affect how power is exercised. The EU is now in a situation where it is forced to

32 *See* E. Wennerström. *The Rule of Law and the European Union*. Iustus. 2007.

set out, in much clearer detail, the meaning of the rule of law in the Copenhagen document, the Treaty of the European Union, and the EU rule of law mechanism. This illustrates the potential downsides of approaching the rule of law primarily based on its constitutive parts rather than as a whole.

So, where does this lead? The ends-based approach to rule of law and automated governance and AI? Two things are worth highlighting. The first is straightforward, which is the issue that the rule of law is not a quick fix for social problems writ large. The rule of law is conservative by nature and serves the purpose of minimising arbitrariness by adding rules to political processes. As Waldron frames it, maybe there is no exemplar rule of law, but just a problem that has 'preoccupied us for 2,500 years: how can we make law rule? On this account, the Rule of Law is a solution-concept, rather than an achievement concept, the concept of a solution to a problem we're not sure how to solve; and rival conceptions are rival proposals for solving it or rival proposals for doing the best we can in this regard given that the problem is insoluble'.[33] The historical record of a problem that is insoluble speaks for perseverance when applying old concepts to new problems, which leads to the second aspect, namely time and changes over time.

Law is not static over time, but it is almost always late. The empirical world moves ahead with or without laws and regulations in place, as we can see with AI and automated decision-making today. Thus, the second aspect to highlight in relation to the ends-based approach to the rule of law is timing. In automated governance and digitalisation generally, it is often suggested that old concepts may not be suited to solve new problems. At a first glance, this seems correct, considering the complexity of technology and the multitude of actors and agencies involved. But this is a perspective based on checklists, not on old concepts failing the task of showing their relevance in relation to new problems. Instead, arguably, it is that old prescriptions (based on old concepts) perform sub-optimally when confronted with new problems. This is not only an issue of timing, but also one of varying institutions and frameworks: 'The institutions that occurred to Aristotle to distinguish "the rule of law" from "that of any individual" were not those specified in Magna Carta;

33 J. Waldron, *Is the Rule of Law an Essentially Contested Concept (in Florida?).* 21 Law & Philosophy 137, p. 158. 2002.

those had little in common with the ones that drew Montesquieu; what he lit upon was different from Dicey's homegrown selection; these from Hayek's; those from Oakshott's; his from Fuller's; his (though not so much) from Raz's; any of theirs from Waldron's; his from those chosen by rule of law indexers; theirs from each other.'[34] The very old response to the perennial problem of social ordering and arbitrary use of power is therefore not out of sync, but the specific solutions that grow out of the idea at any given time might very well be.

In the following, two core safeguards of the rule of law are examined – legality and accountability – specifically for their relevance to automated governance and the point of minimising arbitrary power.

3 Law should rule, but how?

Automated governance is most straightforward and least problematic when it concerns binary decision-making that can be quantified (for example infrastructure charges). A greater challenge is if the law is vague or lacks precision, which creates a specific complexity in automated decision-making (and challenge to suitable safeguards). Translating broad or unspecific laws, rules and regulations into code entails a risk of supplementing codes with data and automated regulation, rather than regulation through automation.[35]

It is not clear how a transfer to automated governance that includes machine learning should deal with situations where the law leaves scope for assessment. Should the assessment be tied to algorithms for machine learning, and thereby substitute human discretion for automated discretion? It seems important that a legal framework for control and appeal of decisions, handled by humans, is put in place as a safeguard and for handling potential risks and legal losses for the individual. From the perspective of legality, which is a cornerstone of the rule of law, automation can have serious implications. Not least considering the rapid development of machine learning and

34 Krygier, *supra* note 11, p. 750.

35 Juridik som stöd för förvaltningens digitalisering, SOU 2018:25, p. 158. This could potentially pose a constitutional problem in Sweden, specifically regarding rules on legislative power if a program used for automation in some cases could be a regulation or rule in itself.

the expectation that this will form a common part of governance in a near future.

The discussion in Sweden, but also in other countries, treats automation from a legal technical point of view. For instance, it is unclear to what extent Swedish law allows for an automation of public decision-making, or if a clarification of the law is needed.[36] It is also unclear how legal certainty can be upheld in automated decision-making and how privacy and information security can be protected.[37] The legal technical lens might be an effect of the cost-effectiveness considerations underlying many automation initiatives.[38] Through recent government inquiries in Sweden and a new Agency for Digital Government, the government is trying to catch up. However, there is no coherent and long-term responsibility for analysing and following up on the politics of digital development.[39]

The recent Administrative Procedures Act from 2018 is an example where the catching up with technical developments only reaches a certain level of safeguarding. The new law includes a general paragraph allowing automated decision-making, something which was considered necessary to reflect recent developments at public agencies, and to avoid having to issue specific regulations for each agency or area of public administration. This allows for a threshold of legality when it comes to automation, but the law and the preparatory work do not fully include suitable safeguards, as stipulated in the EU's data protection regulation and in the guidelines from the Article 29 Working Party or the HLEG-AI on trustworthy AI. The preparatory works for the Administrative Procedures Act also completely ignore the issue of accountability when it comes to what body should be seen as responsible, and held accountable, in auto-

36 A government inquiry examining how automated decision-making can be implemented at the level of municipalities and regions is under way at the time of writing this paper.

37 *See* the discussion in Government of Sweden, *Automatiserade beslut – färre regler ger tydligare reglering,* SOU 2014:75. 2014, and criticism from the Parliamentary Ombudsmen, for example in an inspection of the Transport Agency, *Inspektion av Transportstyrelsen, Körkortsenheten i Örebro, den 26-28 oktober 2011*, dnr 4728-2011.

38 A report from the Swedish Agency for Digital Government cites cost savings of six percent of all public expenditures through the introduction of AI.

39 Statskontoret. *Fortsatta former för digitaliseringspolitiken. Utvärdering av Digitaliseringsrådet och kartläggning av regeringens styrning*. 2020:3:7.

mated governance: that which designed the algorithm or the public agency using it for decision-making.[40]

The GDPR Art. 22.1 generally prohibits decision-making based solely on automation, including profiling, if it produces legal effects or affects an individual in a similar way. There are exceptions for public authorities when automation is authorised by Union or Member State law, which lays down suitable safeguarding measures. The Article 29 Working Party mentioned that safeguards could mean specific information to the data subject, a right to obtain human intervention, to be heard and to get an explanation of the decision, or a right to challenge the decision.

In Sweden, the legal basis presents some challenges and legal issues regarding automation at the municipal level remain. Legality, however, is more than just a sharp line against *ultra vires* decisions. Legality also includes that the legal basis should be (fairly) clear, understandable, and transparent. This is in line with HLEG-AI's ethics guidelines, where transparency is a cornerstone for trustworthy AI and the European Commission's white paper on AI.[41] Automated governance with current technology and certainly in the future, if machine learning is more frequently employed, might mean that the grounds for a decision would be harder for an individual to decipher. It is no longer 'just' a public servant applying laws and following guidelines for reaching a decision, but a program working based on pre-determined parameters.

Pursuant to GDPR, individuals have a right to be informed when a decision is based solely on automation. How far the right to be informed extends is debated among lawyers and legal scholars, but the guidelines from the Article 29 Working Party on automated decision-making (now endorsed by the European Data Protection Board) suggest that the right to be informed should encompass general information about the logic involved and the significance and potential consequences of the processing. The Article 29 Working Party acknowledges that the logic involved, due to the growth and complexity of machine learning, 'can make it challenging to understand how an automated decision-making process or profiling

40 Karlsson, *supra* note 6, p. 76.

41 High-Level Expert Group on AI, *Ethics Guidelines for Trustworthy AI*. European Commission. 2019; European Commission, *White Paper. On Artificial intelligence – A European approach to excellence and trust*. COM(2020) 65 final. 2020.

works'.[42] Thus, the right to be informed seems to encompass general information that is comprehensive, so the data subject can understand the reasons for the decision, but does not encompass an explanation of the algorithms used or disclosure of the full algorithms.[43]

Even if the right to be informed should be viewed more extensively, including a detailed explanation of the algorithms used, what use is that information to an individual? How can this information be properly assessed and, if desired, acted upon in the sense of making an informed choice on what to do? The Article 29 Working Party Guidelines propose that graphics, visual techniques, and other tools be used for explaining the process of an automated decision and its consequences.

This leads back to the sociolegal and behavioural perspective mentioned earlier – that we understand law through metaphors and embodiments. When the human element is removed from public decision-making, and when algorithms are of such complexity that explaining their function would be pointless in aiding understanding of how a decision was made, other means are necessary to support a cognitive anchoring of what automated decisions mean.

4 Accountability – Who guards the guardians?

Accountability is a central feature of the rule of law and its ability to minimise arbitrary power. Accountability pre-supposes an existing responsibility, since it is often impossible to separate the two (and when it is done, problems tend to arise). Automated decision-making, where machine learning or artificial intelligence is employed, generates several challenges to accountability ground rules as we generally understand them. Here, accountability means 'the implicit or explicit expectation that one may be called on to justify one's beliefs, feelings, and actions to others'.[44] The ground rules of accountability which specify 'who must answer to whom, and for what, are essen-

42 Article 29 Data Protection Working Party, *supra* note 4, p. 14.

43 *Id.*

44 J. S. Lerner & P. Tetlock, *Accounting for the Effects of Accountability*. 125 Psychological Bulletin, p. 255. 1999. *See also* M. B. Scott & S. Lyman, *Accounts*. 33 American Sociological Review, pp. 46–62. 1968.

tial features of human social life'[45] and as such have been the core ingredient in the evolution of the rule of law.[46]

It is easy to call for more accountability where public decision-making is concerned. However, accountability is not an unalloyed good, since how it is exercised (the context), in relation to what (i.e., process or outcomes) and when (timing), influences whether it has a positive effect on judgments and decision-making.[47] These considerations, it is argued, matter even more for automated governance, in particular regarding AI – where the identification of accountability is more complicated due to context, timing and process or outcome factors. In addition, it also matters how responsibility is framed for someone to perceive themselves to be accountable. There must be a legally founded scope for choice regarding how to act or what decisions to make. Unrealistic expectations or responsibility held back by powerful constraints, so that certain tasks cannot be performed, should not generate accountability.

Where automated governance is employed for handling quantifiable and easy decisions, the question of accountability deviates little from how it is typically intellectualised. In situations where the law leaves scope for assessment and discretion, or situations where machine learning is employed, accountability becomes more difficult. It becomes difficult because technology does not simply serve as a tool, but is actually replacing human action.

Timing also carries weight for accountability having a positive effect on decision-making. Research suggests that post-decision accountability – introducing accountability only after a decision has been made – seems to strengthen commitment to earlier courses of action. Pre-decisional accountability, where someone knows they will be held accountable before they make a decision, lessens commitment to a specific course of action. Studies have shown that participants in post-decisional situations would 'think of as many

45 W. Chang, et al., *Accountability and adaptive performance under uncertainty: A long-term view*. 12 Judgment and Decision Making, p. 610. 2017.

46 A. Sajó & R. Uítz. *The Constitution of Freedom: An Introduction to Legal Constitutionalism*, p. 306. Oxford University Press. 2017.

47 J. S. Lerner & P. E. Tetlock, 'Bridging individual, interpersonal and institutional approaches to judgement and choice: The impact of accountability on cognitive biases' in *Emerging Perspectives in Judgments and Decision Making*. S. Schneider & J. Shanteau (eds.), Cambridge University Press. 2002.

reasons as they could to bolster their decision'. However, where participants learned of their need to justify their decisions before they formed an opinion, this seemed to cause them to 'impartially consider whether or not to continue their commitment'.[48] There are more elements to this, however, as Lerner and Tetlock show in their summary of the literature on judgment and decision-making: '…the framework predicts that integratively complex and open-minded thought is most likely to be activated when decision makers learn prior to forming their opinions that they will be accountable to an audience (a) whose views are unknown, (b) who is interested in accuracy, (c) who is reasonably well-informed, and (d) who has a legitimate reason for inquiries into the reasons behind participants' judgment choices.'[49]

Tetlock's and Lerner's framework suggests several problems regarding AI in relation to accountability. For one thing, the locus of accountability shifts whenever technology is employed, and arguably becomes more difficult to identify as technology for dealing with intricate problems becomes more complex. It transfers accountability from a place where we typically find accountability mechanisms in relation to public servants, to developers, programmers and those procuring technical services for use in the public domain.

Similarly, the place of control and oversight also shifts. If public agencies rely on more automated systems for their decision-making, then control – in the sense of understanding, having the ability to monitor and to correct or adjust errors – moves from public agencies to private companies. Accountability and the ability to exercise a proper system for control and oversight is also affected in a temporal way, not just in the physical embodiment of who is responsible for what. Several municipalities in Sweden have begun using automated decision-making for income support cases. The legality of doing so is not clear, but a larger issue relates to accountability. To what extent is the municipality of Trelleborg in southern Sweden, where automated decision-making is already in use, accountable for decisions made by algorithms designed elsewhere, outside their immediate control? Here, the timing issue comes into play, according to Tetlock's and Lerner's account.

48 *Id.* p. 14.

49 *Id.* p. 15.

To what extent was the design phase of the algorithms used by Trelleborg's municipality subject to accountability claims – was there a pre-decisional accountability in the sense that the developer took care to construct a program through an 'integratively complex and open-minded thought'? Or is it the public servants at the municipality who should (legally) bear the responsibility for any faulty decisions? If so, what does this post-decisional accountability mean in terms of the public servant's ability to correctly and thoroughly investigate the decision-making process in an objective and impartial manner? It is worth noting that the proposed regulation from the European Commission attempts to cover this complexity by shifting the focus to encompass the AI supply chain and the responsibilities of suppliers and users of AI.[50]

The literature on judgment and decision-making suggests that post-accountability mechanisms reinforce a commitment to earlier courses of actions. From a public servant point of view, much like the individual or client perspective, you are operating a system that is complex and difficult to understand and survey, producing decisions for which you must rely on the correctness of a previous course of action.

An unintended side effect of AI systems, where accountability is difficult to locate (who, where) or to fix in time (when), is that they may produce a governance situation where large numbers of civil servants are responsible for certain areas (income support, taxation, infrastructure charges, employment benefits), but they are not accountable. The decisions produced by algorithms that public servants then supervise will create only a nominal sense of accountability. However, from the perspective of individual citizens, users or clients, public servants may very well be responsible and accountable for automated decisions, creating inconsistencies in the governance framework that disrupt the link between decision-makers on the one hand and social systems on the other. This is the narrative of the legal risks typically discussed in relation to automation, and the sociolegal risks of creating a dissonance in the system of public rule.

50 European Commission, *Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonized Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain European Union Acts*, Brussels, 21.4.2021 COM(2021) 206 final, available at https://op.europa.eu/en/publication-detail/-/publication/e0649735-a372-11eb-9585-01aa75ed71a1/language-en/format-PDF.

5 Robustness of the rule of law

This paper argues that the rule of law, as a point of analysis, has a lot to offer the process of digitalisation and automation of governance. The proposed analysis is cumulative in the sense that the rule of law is constituted by a set of principles or safeguards where the whole is greater than the sum of its parts.

The argument here is for robustness in a systematic way of dealing with how power is exercised. This includes the capability to handle risks and respond to unintended negative effects with a sound institutional structure. This is a challenge in the Swedish context, and likely also in other countries with decentralised governance systems. The embedded individual autonomy of public agencies adds an extra layer of complexity in the Swedish case.

The standards expressed in GDPR and in the discussions in the HLEG-AI are important for setting out specific safeguards in the institutional framework of member states. The safeguards described in the GDPR preamble require that when automated decision-making is used: 'specific information to the data subject and the right to obtain human intervention, to express his or her point of view, to obtain an explanation of the decision reached after such assessment and to challenge the decision. Such measures should not concern a child.'[51] In the earlier proposed regulation from the European Parliament on AI, the following were suggested as conditions for public authorities: 'Notes that the development, deployment and use of artificial intelligence, robotics and related technologies, by public authorities are often outsourced to private parties; considers that this should not compromise the protection of public values and fundamental rights in any way; considers that public procurement terms and conditions should reflect ethical standards imposed on public authorities when applicable.'[52] In the European Parliament and Council Artificial Intelligence Act, a similar but a somewhat watered-down approach was taken, stating that public authorities

51 European Union, Council Regulation 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC, OJ 2016 L 119/1 Recital 71.

52 European Parliament, *Framework of ethical aspects of artificial intelligence, robotics and related technologies*. P_9TA(2020)0275. Art. 77.

which put into service high-risk AI systems may 'adopt and implement the rules for the quality management system as part of the quality management system adopted at a national or regional level […]'.[53]

Far from these sublime preambles and principles, it is in the mundane reality that automated governance is taking place, at the level of municipalities and regions. Here, the robustness of safeguards will be problematic – and already is. The Swedish Agency for Digital Development's mapping report on AI identified several challenges that public agencies face with their architecture for AI solutions. They include aspects such as uneven AI competencies within public agencies, uncertainties on how to handle ethical and legal aspects, and the difficulty of public sector management to adequately respond to changes caused by automated decision-making.[54]

The Swedish Agency for Public Management has a similar description of constraints when it comes to sufficient resources and competencies to, for example, secure satisfactory procurement processes regarding AI. The Agency for Public Management raises another point, namely the shift in 'culture' at public agencies that follows from a greater reliance on automated decision-making. Public servants with IT competence will become more influential in the day-to-day work of public authorities, and many in this professional group are short-term problem-solvers or external consultants.[55] Common to this professional group is that public 'ethics' and public service culture are not dominant characteristics. Relying on professional groups outside the public sector has proved difficult in terms of maintaining a culture of good governance.[56]

Digitalisation of governance is moving forward in many places at once. In Sweden, there is no coherent national definition of AI and no clear framework for how to handle automation. Several municipalities are rolling out automation of decision-making, though the

53 European Commission, *Proposed regulation from the European Parliament and of the Council. Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Act*s 2021/0106(COD).

54 Agency for Digital Government (DIGG), *Främja den offentliga förvaltningens förmåga att använda AI,* p. 28 f. 2019.

55 Statskontoret, *Förvaltningspolitik i förändring – långsiktiga utvecklingstendenser och strategiska utvecklingsbehov*, p. 46. 2019.

56 Statskontoret, *Att göra eller köpa? Om outsourcing av statlig kärnverksamhet*. 2015.

legal basis for using automated processes is unclear (a government inquiry is expected to report its results on automation and municipal governance in 2021).

Beyond the technical discussions, which involve legal technical responses, lies a deeper concern. Law and the rule of law are more than just rules. It is also an understanding of how the rules are 'put together, how the system is structured, how the rules are interpreted'.[57] How we cognitively understand law and the metaphors and embodiments we use to make sense of the abstract also extend to the legitimacy and the weight afforded to letting law rule.

One's imagination need not stretch far to suggest that an automation of governance and machine learning where code replaces rules can be at odds with how law has been understood for a very long time. There is a traditionality about law which largely goes unnoticed. Krygier has shown how law, as tradition, has three characteristics: a pastness, an authoritative presence and a process for handing over the law between generations. 'Such understanding is important not merely, or even especially, for the historian or theorist of law who seeks to account for these things. It is in fact an unsung precondition of practical lawyering; the "tacit" knowledge which underlies competence within any legal or indeed any social practice.'[58] Moreover, Krygier writes, law 'rests upon mountains of inherited tradition, preserved, referred and deferred to by highly developed institutions and practices of tradition maintenance'.[59] This is also a tradition that is inherently linked to human agency. The making of law and the application of law are things humans do, going back to Hammurabi and Draco and ever since Aristotle suggested that man-made laws should rule over men.

Highly developed institutions may well, in a near future, be machine learning institutions. This should give rise to in-depth discussions regarding what this means for the traditionality of law, and the tradition of law as something established and understood through metaphors and embodiments, as politicians, public officials and programmers attempt to digitalise governance. Unintended consequences may be that more public servants are responsible,

57 A. Watson, *The Making of the Civil Law* at p. 14. Harvard University Press. 1981.

58 M. Krygier, *Law as Tradition*. 5 Law and Philosophy, p. 246. 1986.

59 *Id.,* p. 256.

but not accountable, and that the culture of civil services and the traditions expressed in ethics, practices and understandings shift to something different.

6 Conclusions

An all-too common metaphor for technology like AI is to describe it as something magic, sometimes as dark magic – and magic is an unsound basis for governance. In what is almost a paraphrase of Max Weber: the best antidote to charismatic, magical, and transcendental governance is form, process, and structure. Thus, despite the magical undertones of complex technology, it is also heralded as an antidote to human fallibility, with bias, laxity and nastiness contrasted against regularity, precision and logic.

Technology is however (or may well be) magic in the sense that to all but a few, the way in which automated decision-making by algorithms works is a deep mystery. The mystery is moreover hidden in plain sight, because we increasingly use and depend on technology in all aspects of life, without understanding the fundamentals (the combustion engine, for instance, was and is far easier to understand than face recognition software). Thus, an underlying risk from everyday use of a complex technology we do not fully understand is that it might become something that is, as Wittgenstein suggested, 'hidden to us because of [its] familiarity'.[60] Thus, decisions based on a smart design are taken for granted, not because they emanate from a machine with less fallibility than humans, but because we are familiar with technology setting the parameters within which we live our lives.

A balanced discussion on AI and automated decision-making is important. Criticism of AI and automation sometimes gives the impression that governance today is inoculated against discrimination, repressive law, inequality, and malice. It is not, and the relevance of the rule of law over time is a testament to this fact. Some assessments of automated decision-making point to the benefits in terms of removing the risk of arbitrariness that results from human handling. While there are strong arguments for minimising arbitrariness through automated decision-making, it should be clear that

60 L. Wittgenstein, *Philosophical Investigations* at 50. Oxford University Press. 1967.

this really means that the risk of arbitrariness is shifted from the human decision-making process to the human design of algorithms.

Digitalisation of the public sector is a political priority across Europe. In Sweden, with the recent establishment of an Agency for Digital Government, digitalisation is sometimes construed as a project. The ground rules of a project are that it has a start date and an end date, with a set of implementable activities planned between these dates, to reach specific outcomes. It is important to recognise that there is no end date when it comes to digitalisation. The question therefore becomes not one of *doing* or *reforming* something, i.e., revising laws to allow for automation or procuring software to handle large amounts of case law, but of safeguarding values and ethics in the public sector in a sustainable manner. Of doing, as Waldron suggested, the best we can, given that the problem is insoluble.

Between Risk Management and Proportionality: The Risk-Based Approach in the EU's Artificial Intelligence Act Proposal*

TOBIAS MAHLER

1 Introduction

The European Commission has issued a proposal for an Artificial Intelligence (AI) Regulation[1] (hereinafter 'Proposal' or Artificial Intelligence Act, AIA), laying down harmonised rules concerning certain AI systems in the European Union (EU). A key regulatory approach in the Proposal is that the development and use of AI systems is regulated based on risk level. The Proposal's 'risk-based approach' consists of the use of risk levels as thresholds for specific requirements in the Proposal. AI systems that represent unacceptable risks are prohibited and high-risk systems must comply with specific requirements. Less risky systems must comply with fewer or no requirements.

* This research was partly supported by the project 'Vulnerability in the Robot Society' (VIROS, grant number 144789) financed by the Research Council of Norway. The author wishes to thank the members of the VIROS project team, especially Lee Bygrave, Rebecca Schmidt, Mona Naomi Lintvedt and Live Sunniva Hjort, as well as Samson Esayas for useful comments on an earlier version of this paper.

1 Commission Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts, COM (2021) 206 final (Apr. 21, 2021) (hereinafter 'Proposal').

The Proposal explicitly aims to manage the risks of AI systems employed in the EU, so risks are its main object and justification.[2] Therefore, the Proposal emphasises the need to establish rules that are proportionate and effective.[3] To achieve this proportionality, the risk-based approach utilises risk levels (e.g., 'high risk') to trigger requirements for AI systems. Thus, key parts of the Proposal merge risk thinking with rulemaking. The word 'risk' occurs 344 times in the Proposal and many more times in the accompanying Explanatory Memorandum, Annexes and Impact Assessment. Risk is also emphasised in some of the literature focussing on the regulation of AI.[4] Therefore, it comes as no surprise that many elements of the Proposal are in some sense 'risk-based'. Nevertheless, only one is identified as the 'clearly defined risk-based approach'. The purpose of this contribution is to distinguish the so-called 'clearly defined risk-based approach' from the other risk-based approaches used in the Proposal and to unpack whether the concept of risk is applied in the Proposal in a coherent, logical and consistent way.

The 'clearly defined' risk-based approach raises questions about its aim, logic and limitations. What exactly characterises the approach? Is this an example of the European lawmaker engaging in a formalised risk management process by identifying, analysing and treating risk? Indeed, at some level, this seems to be the case: the EU identifies AI risks as a regulatory concern, distinguishing various risk levels and proposing law to manage these risks. This could be seen as the lawmaker's attempt to act more rationally in that it employs a rigorous risk management approach. However, on closer examination, there are indications that the risk-based approach is not as rigorous as it might initially appear. Ultimately, this paper considers what problem, if any, the risk-based approach seeks to solve. It suggests that the problem to be solved by the approach is not primarily how to manage AI risks, but how to avoid a potentially over-broad scope of the regulation—a potential created by the broad definition of AI

2 *Id.*, Recital 4.

3 *Id.*, Recital 14.

4 See, e.g., Michael Guihot, Anne Matthew and Nicolas Suzor, *Nudging Robots: Innovative Solutions to Regulate Artificial Intelligence*, 20 Vanderbilt J. Ent. & Tech. L. 385, 426 (2017).

included in the Proposal.[5] An alternative to applying it would have been a blanket regulation of all AI, which might have introduced excessive obligations on AI producers and users, disproportionately hampering the development of societally desired and economically lucrative AI. Paradoxically, the risk-based approach's aim and utility are not primarily to manage risk but instead to ensure legislative proportionality.

The paper primarily aims to analyse the Proposal, but in doing so, it also introduces, presents and describes part of the Proposal, as not all readers will have studied it in detail. Moreover, the law-making process may move on from where it is at the time of writing, so it may be useful to document some key features of the current Proposal, which forms the starting point of this paper.

The remainder of the paper is structured as follows: Section 2 commences by framing the main risk-based approach from a risk management perspective. Section 3 elucidates that the risk-based approach is only one of several risk-focussing legislative techniques included in the Regulation, as risk management thinking has influenced large parts of the Proposal. Subsequently, Section 4 discusses the concept of 'risk' included in the Proposal. Parts of the Proposal require a qualitative approach to risk rather than a quantitative one characterised by risk calculations. Section 5 addresses the potential rigour that the risk-based approach could contribute to the law-making process. The risk-based approach is presented as a seemingly rigorous and rational methodology, tailoring the rules to the 'intensity and scope of the risks that AI systems can generate'.[6] Section 6 returns to the risk-based approach's main function, ensuring the Proposal's proportionality rather than managing risk.

5 See also, Sebastian Felix Schwemer, Letizia Tomada and Tommaso Pasini, *Legal AI Systems in the EU's Proposed Artificial Intelligence Act*, Proceedings of the Second International Workshop on AI and Intelligent Assistance for Legal Professionals in the Digital Workplace (2021).

6 Proposal, *supra* note 1, Recital 14.

2 The main risk-based approach from a risk management perspective

Recital 14 describes the clearly defined risk-based approach as follows:

> In order to introduce a proportionate and effective set of binding rules for AI systems, a clearly defined risk-based approach should be followed. That approach should tailor the type and content of such rules to the intensity and scope of the risks that AI systems can generate. It is therefore necessary to prohibit certain artificial intelligence practices, to lay down requirements for high-risk AI systems and obligations for the relevant operators, and to lay down transparency obligations for certain AI systems.[7]

The risk-based approach builds on both the 2020 AI White Paper,[8] which laid out the foundations of a future regulatory framework for AI in the Union, and a resolution passed by the European Parliament.[9] Since its drafting, the approach has evolved from a dichotomy between 'high-risk AI' and other AI to a more elaborate AI risk categorisation. The Proposal addresses various categories of AI systems based on risk levels indicating the magnitude of risk, as described below.

First, certain AI practices are considered a concern so significant that they are prohibited under Article 5. The Proposal does not explicitly assign a risk level to these practices, but from a risk management perspective, one could say that these risks are unacceptable—arguably because of their excessive risk level. For example, Article 5 prohibits an AI practice that exploits any vulnerabilities of a specific group of persons because of their disability, if further conditions are fulfilled.[10]

7 *Id.*

8 *White Paper on Artificial Intelligence – A European Approach to Excellence and Trust* 16, European Commission (February 19, 2020), available at http://ec.europa.eu/info/sites/default/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf (last accessed 12 September 2021).

9 *European Parliament Resolution of 20 October 2020 with Recommendations to the Commission on a Framework of Ethical Aspects of Artificial Intelligence, Robotics and Related Technologies* (2020/2012(INL)), European Parliament (2020).

10 Proposal, *supra* note 1, Art. 5 (1) (b).

One level below these unacceptable practices are 'high-risk' AI systems, which are the focus of the Proposal. Article 6 defines what is considered high-risk AI; this definition is based on Annex III. Article 7 offers a mechanism for updating the catalogue of high-risk systems contained in the Annex. If an AI system is classified as 'high-risk', it triggers detailed requirements,[11] with specific obligations for various parties.[12] It is beyond the scope of this paper to present those rules. However, it is worth noting that the Proposal includes a requirement to have a risk management system pursuant to Article 9. Thus, if an AI system is considered high-risk AI, it triggers an obligation to manage the risks of said system. Thus, risks are identified and managed at several levels of abstraction, as discussed further in Section 3.

A third level focusses on AI systems, such as chatbots, intended to interact with natural persons and AI systems used to generate or manipulate image, audio or video content.[13] These systems are not classified as 'high-risk', and the Proposal fails to assign any other explicit risk level. For these systems, the Proposal only foresees transparency rules to ensure that humans are informed that they are interacting with AI. It follows from the logic of the Proposal that, given the more limited obligations, the risk level of these systems must be lower than that of 'high-risk' systems. Indeed, in the fact sheet released by the Commission in conjunction with the Proposal, this group is denominated 'limited risk'.[14]

If we follow this line of thinking—sorting risks based on magnitude—the remaining AI systems represent the lowest risks. The Proposal does not assign these systems to any named category or risk level, but the Commission nevertheless refers to them as 'minimal risk' systems.[15] They arguably constitute the largest group of AI systems in practice. Examples include applications such as simple image recognition systems and email spam filters, which do not raise the same concerns as high-risk AI systems do. Many different

11 *Id.*, Title III, Ch. 2.

12 *Id.*, Ch. 3.

13 *Id.*, Art. 52.

14 *EU Fact Sheet: Excellence and Trust in Artificial Intelligence*, European Commission (2020), available at https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/excellence-trust-artificial-intelligence_en (last accessed 12 September 2021).

15 *Id.*

technologies qualify within the Proposal's broad definition of AI systems, but are not included in any of the above risk levels, remaining unregulated by the Proposal. This does not exclude the possibility that these systems are regulated by other legal frameworks[16] and the possibility of soft law codes of conduct also remains.[17] Still, the Proposal puts the lowest regulatory burden on the largest group of AI systems.

In summary, the risk-based approach focusses on organising AI practices and systems based on risk level. The resulting classification triggers requirements and obligations, including the duty to manage the risk of AI systems. By classifying different types of AI systems by risk level, the lawmaker appears to focus on managing AI risks; it seems to take on the role of a risk manager. In international standards, 'risk management' is used as a technical term that refers to a set of coordinated activities to direct and control an organisation with regard to 'risk' of a nature to be specified.[18] The organisation carrying out the risk management could here be the EU lawmaker, and the risk in question could be called 'AI risk'.

According to the International Organization for Standardization (ISO) standard 31000, risk management consists of one or more risk assessments. The term 'risk assessment' refers to the overall process of risk identification, risk analysis and risk evaluation.[19] For example, a risk assessment could focus on the maiden voyage of the Titanic. The assessment would commence by identifying risks and describing them, for example in terms of an event (the ship collides with an iceberg) and its consequences. Once a risk is identified, it can be

16 See, e.g., European Union, Council Regulation 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC, OJ 2016 L 119/1 [hereinafter General Data Protection Regulation ("GDPR")]; cf. Lee A Bygrave, *Machine Learning, Cognitive Sovereignty and Data Protection Rights with Respect to Automated Decisions*, University of Oslo Faculty of Law Research Paper No. 2020-35 (2020).

17 Proposal, *supra* note 1, Title IX.

18 *ISO Guide 73:2009: Risk Management – Vocabulary*, s. 3.2, ISO (2009), available at https://www.iso.org/standard/44651.html (last accessed 12 September 2021).

19 *Id. at* s 3.3.3.

analysed to estimate the risk level.[20] This *risk analysis* serves to create understanding of each risk and determine the risk level by combining the estimated likelihood of the event with the extent of the consequences.[21] Risk criteria can be used to assess the significance of a risk. These criteria can include the organisation's risk appetite, as well as other factors (e.g., legislation). As a result, some risks can be accepted, while others must be addressed. A risk assessment of the *Titanic* is concrete compared with the abstract, broad and ambiguous project of assessing AI risks in the EU, particularly when many risks associated with new or emerging AI technologies are hard to anticipate or foresee.[22] Accordingly, the legislative risk assessment presented in the AIA Proposal is a challenging endeavour.

Large parts of the AIA Proposal are phrased in the language of risk management. For example, according to the Explanatory Memorandum, risks should be 'calculated taking into account the impact on rights and safety'.[23] Furthermore, the Proposal aims to tailor the rules to the 'intensity and scope of the risks that AI systems can generate'.[24] What is meant by 'risk intensity' remains unclear—at least if one reads the Proposal with risk management terminology in mind; most likely it refers to the risk level, i.e., the magnitude of risk. This framing supports the impression that the risk-based approach is essentially an attempt to apply risk management to a legislative intervention based on a putatively rigorous methodology and risk criteria. The 'clearly defined risk-based approach' is strongly emphasised in the Proposal, but it is only one of several regulatory techniques to manage the overall risk of AI systems in the European Union. Despite the rhetoric around this risk-based approach, the role of risk in the Proposal is highly multi-faceted.

20 According to the ISO, the level of risk is the magnitude of a risk or combination of risks, expressed in terms of the combination of consequences and likelihood; see *id.*, s 3.6.1.8.

21 The ISO defines a consequence as the outcome of an event affecting objectives, see *id.*

22 Marjolein van Asselt, Ellen Vos and Tessa Fox, '*Regulating Technologies and the Uncertainty Paradox*' in M.E.A. Goodwin, E.J. Koops and R.E. Leenes (eds), Dimensions of Technology Regulation (Wolf Legal Publishers 2010).

23 Proposal, *supra* note 1, Explanatory Memorandum to the Proposal 8.

24 *Id.*, Recital 14.

3 Other risk-focussing legislative techniques

The AIA Proposal employs the 'clearly defined risk-based approach', in addition to several other elements that could also be labelled 'risk-based approaches'. To avoid confusion, these other approaches are here called 'risk-focussing techniques'. The Proposal contains four types of risk assessments that are performed by various actors and with different levels of concreteness, as summarised in Table 1.

Table 1 Risk-focussing legislative techniques applied in the AIA Proposal.

Approach	Risk	Focus	Provision
Legislative risk assessment	Abstract AI risk	High-risk AI systems	Articles 6, 7
Duty to manage risk	Concrete AI risk	Managing risk	Article 9
Enforcement risk analysis	Concrete AI risk	Risk at national level	Article 65
Compliance incentives	Legal and financial risk	Infringements	Article 71

First, under the main risk-based approach mentioned above and discussed in the remainder of this paper, the legislator classifies *abstract* types of AI practices and systems into a set of risk categories. For example, Annex III classifies certain AI systems used for recruitment as high-risk AI systems, and the underlying risk assessment describes those systems and the risks they create in general terms.[25] Thus, certain broadly described AI use cases are legally qualified as pertaining to a risk level, but in abstract terms only. This could be called *legislative risk assessment,* as it is carried out by the legislator. By comparison, the second risk-focussed technique is much more concrete, with risks identified and managed by a different set of actors. Article 9 necessitates the establishment of a risk management system for high-risk AI systems. It requires a concrete, continuous and iterative risk management process that runs throughout the lifecycle of a given AI system. This differs from the legislative risk assessment mentioned above, where AI risk is assessed once (when classifying the system type) and only abstractly. Pursuant to Arti-

25 *Id.*, Annex III and Annex to the Impact Assessment 43.

cle 9, the AI provider and other actors responsible for complying with this requirement must assess and manage the concrete risks of specific AI systems. This can also involve the classification of risk by levels. Therefore, it is possible that a provider of a 'high-risk' (with hyphen[26]) AI system estimates a specific concrete risk as 'high risk' (without hyphen). In addition, other risk levels—such as 'low risk' or 'very high risk'—are also available. For a hypothetical and admittedly artificial example, consider the possibility that the provider of an AI recruitment system (which falls into the abstract class of 'high-risk AI') identifies the following risk: the speed and simplicity of the concrete recruitment system could render it so popular among applicants that the system might crash, leading to system downtime. Let us further assume that this is unlikely to happen and would have limited consequences. Hence, this specific risk may be considered a 'low risk', but the overall assessment remains: it is a 'high-risk' system. This may be confusing, but would not be a contradiction, because the assessment is carried out at a different level of abstraction and probably employs different criteria for classifying risk. We return to this possibility of confusion later, but first add yet another layer of risk assessment.

Article 65 foresees a procedure for dealing with AI systems presenting a risk identified at the national level by the authorities of a Member State. Again, the focus is on a concrete system and the risks it represents. The market surveillance authority of a Member State assesses whether a system presents 'a risk', but the notion of risk here is different from the abstract concept of 'high-risk' AI systems. Notably, Article 65 does not require the presence of 'high risk'—only 'risk'; however, the risk level will arguably still play a role in triggering a follow-up by the authority because the lowest risks can probably be accepted. Thus, risk is identified and assessed by yet another actor—not the lawmaker or the AI producer, but a national enforcement agency. This implies that the risk levels and respective criteria could again differ from the first two risk assessments mentioned above. The function of this third risk assessment is to trigger further investigations and potentially enforcement actions,

26 Cf. the use of 'high-risk AI system' in the Proposal, e.g., in Art. 6.

the result of which can be a fine up to 6% of an offender's total worldwide annual turnover.[27]

Non-compliance with the requirements of the AIA may ultimately imply a risk for the actors regulated by the Proposal. An administrative fine would imply a financial or legal risk[28] as opposed to an AI risk. Thus, it could be considered a legal risk assessment,[29] a financial risk assessment or a broader enterprise risk assessment.[30] These actors' risk assessments should provide incentives to comply with the requirements of the AIA. The potential imposition of financial risk constitutes another risk-based feature of the Proposal, involving a transformation of the risks in two distinct ways. First, risk is transferred from the stakeholders who might ordinarily be affected by the AI risk (e.g., job applicants discriminated against by an AI recruiting system) to the actors controlling the risk (e.g., AI providers). Second, in this transfer of risk between actors, AI risk is simultaneously substituted by financial risk. Clearly, this imposition of financial risk by law is nothing new, and the fine levels appear to be modelled after the system established by the General Data Protection Regulation (GDPR) for administrative fines.[31] Still, the possibility of fines ultimately increases the risk for AI providers, thus constituting yet another risk-focussing legislative technique.

Although the Proposal is framed as being based on a single risk-based approach, it really combines multiple risk assessments into a relatively complex overall process for managing AI risk. The Proposal's rhetoric around the 'clearly defined risk-based approach' disregards the various other ways in which risk is employed as a regulatory tool in the Proposal. Thus, the utility of the label 'risk-based approach' is questionable, given the multiple approaches or techniques, which all—more or less explicitly—have some founda-

27 Proposal, *supra* note 1, Art. 71.

28 Tobias Mahler, *Defining Legal Risk* in Corporate Contracting Capabilities: Conference Proceedings and Other Writings (University of Joensuu, Department of Law 2008).

29 Tobias Mahler, *Legal Risk Management. Developing and Evaluating Elements of a Method for Proactive Legal Analyses, With a Particular Focus on Contracts* (University of Oslo Faculty of Law, 2010).

30 Terje Aven and Shital Thekdi, Enterprise Risk Management: Advances on Its Foundation and Practice (2019).

31 GDPR, *supra* note 16, Art. 83.

tion in the concept of risk, albeit at different levels of abstraction. Singling out one of these may be useful enough for law-making purposes, especially as the 'clearly defined risk-based approach' is explained in Recital 14. However, from an analytical perspective, the label is ambiguous. At very least, we should acknowledge that there are several risk-based approaches at play.

Here, we first briefly broaden the focus to view the risk-focussing legislative techniques in the context of the literature on risk and regulation; we then return to the main risk-based approach in the Proposal. Different risk-focussed regulatory techniques are available for lawmakers and regulators, but they are not labelled consistently. For example, 'risk-based regulation' has been advocated as a strategic approach to enforcement by regulatory agencies.[32] Moreover, the literature sometimes speaks of 'risk-based regulatory strategies'. These can be defined as 'collections of strategies that at the very least involve the targeting of enforcement resources based on assessments of the risks that a regulated person or firm poses to the regulator's objectives'.[33] This perspective is only partly applicable to the AIA Proposal, because it is still unclear who will be the eventual regulator and what exactly its objectives will be (beyond managing AI risk on society's behalf).

Black argues that risk plays the four following roles in regulation:[34] (i) providing an object of regulation, (ii) justifying regulation, (iii) constituting and framing regulatory organisations and procedures

32 Julia Black, *The Emergence of Risk-Based Regulation and the New Public Risk Management in the United Kingdom*, Public Law 512 (2005); see further, Karen Yeung and Lee A. Bygrave, *Demystifying the Modernized European Data Protection Regime: Cross-Disciplinary Insights from Legal and Regulatory Governance Scholarship*, Regulation & Governance at 11 (2021).

33 Julia Black and Robert Baldwin, *Really Responsive Risk-Based Regulation*, 32 Law & Policy 181 (2010); see also Henry Rothstein et al., *The Risks of Risk-Based Regulation: Insights from the Environmental Policy Domain*, 32 Environ. Int. 1056 (2006).

34 In the 'risk and regulation' literature, the word '*regulation*' does not refer to an EU *Regulation*, such as the proposed AIA (once adopted, the Regulation will become binding and directly applicable in all Member States, see Treaty on the Functioning of the European Union, Art. 288 (2)). Instead, in the literature, the focus is broadly on the *act of regulating*. One influential definition of 'regulation' refers to 'the sustained and focused attempt to alter the behaviour of others according to defined standards and purposes with the intention of producing a broadly identified outcome or outcomes, which may involve mechanisms of standard-setting, information-gathering and behaviour modification'. Julia Black, *Critical Reflections on Regulation*, 27 Australian

and (iv) framing accountability relationships.[35] All of these roles are at play in the AIA Proposal. For example, Article 65 could also be seen as constituting a regulatory procedure, resulting in accountability (and legal/financial risk) for the AI provider.

The AIA Proposal's emphasis on risk appears to be part of a broader trend towards risk-focussing legislative techniques employed by the EU lawmaker, at least in the context of information and communication technology law. Risks already play a key role in various regulatory contexts, such as data protection law,[36] cybersecurity[37] and product safety.[38] Risks can justify legislative interventions, and many laws include obligations to manage risk in a specific context and an approach related to meta-regulation.[39] In parallel with the Proposal, the EU is advancing regulatory approaches with a strong focus on risk in other contexts, such as the Digital Services Act and the Machinery Regulation. While several of these contain risk management obligations and the GDPR uses 'high risk' as a rule trigger,[40] the AIA Proposal goes one step further in that it employs the 'clearly defined risk-based approach'.

J. Legal Philosophy 1, 26 (2002). In this broad understanding, legislation is merely one instrument in the toolbox of regulation.

35 Julia Black, The Role of Risk in Regulatory Processes (2010).

36 See, e.g., GDPR, *supra* note 16, Arts. 32 and 35.

37 See, e.g., *Directive (EU) 2016/1148 of the European Parliament and of the Council of 6 July 2016 concerning measures for a high common level of security of network and information systems across the Union*, Arts. 14, 1–30, OJ L 194 (2016).

38 *Directive 2001/95/EC of the European Parliament and of the Council of 3 December 2001 on general product safety*, Arts. 1, 4–17, OJ L 11 (2002).

39 Sharon Gilad, *It Runs in the Family: Meta-Regulation and Its Siblings*, 4 Regulation & Governance 485 (2010).

40 GDPR, art 35 (1): 'Where a type of processing in particular using new technologies, and taking into account the nature, scope, context and purposes of the processing, is likely to result in a high risk to the rights and freedoms of natural persons, the controller shall, prior to the processing, carry out an assessment of the impact of the envisaged processing operations on the protection of personal data'. See also Raphaël Gellert, The Risk-Based Approach to Data Protection (2020); Yeung and Bygrave, *supra* note 32.

4 The Proposal's concept of risk

This section discusses the concept of risk employed in the Proposal, especially in the context of the clearly defined risk-based approach. According to the Explanatory Memorandum, risks should be 'calculated taking into account the impact on rights and safety'.[41] When the Explanatory Memorandum uses the expression 'calculating risk', this may be taken to indicate a realist perspective on risk, which is typical of many technical contexts. When a technical expert is called on to assess risk, we expect objective expertise about the risks that exist in the real world. Such a risk assessment could focus, for example, on a machine that may cause bodily harm if it malfunctions. In this case, risk appears to be an aspect of reality that can be objectively assessed and calculated. By comparison, normative considerations regarding rights protection become less prominent as their calculation is more challenging. In contrast, risk management sometimes employs qualitative rather than quantitative approaches. When drafting the Proposal, despite the claims of calculating risks, the Commission probably had such qualitative approaches in mind, rather than mathematical calculations.

In the classic technical perspective, risk is seen as an almost objective calculation based on the properties of the system under analysis.[42] For example, machines may imply technical risks, so they are encompassed by harmonised safety legislation in the EU. A technical expert can assess the risks of a concrete machine and document them in the technical certification. In this sense, risk can be perceived as something outside the law, but its management is regulated by law. This *realist* position dominates in a range of domains, including actuarial applications (in insurance), toxicological and epidemiological research and engineering.[43]

The extreme realist position is reflected in the classic risk terminology in technical disciplines that is visible, for example, in the earlier version (dated 2002) of the ISO risk management vocabu-

41 Proposal, *supra* note 1, Explanatory Memorandum to the Proposal 8.

42 Ortwin Renn, Risk Governance: Coping with Uncertainty in a Complex World 13 (2008).

43 *Directive 2014/33/EU of the European Parliament and of the Council of 26 February 2014 on the harmonisation of the laws of the Member States relating to lifts and safety components for lifts*, Annex IV, A.3, OJ L 96, 29 (2014).

lary. In this version, 'risk' was understood as the combination of the probability of an event and its consequences.[44] From this perspective, assessing risk is a question of calculation, and the risk level is *calculated* based on probability and consequence. In this sense, risk refers to an 'entity, which has an objective existence and is objectively accessible'.[45] Such calculations work best under the assumption that the future conditions of the relevant context are comparable to past conditions, which may be questionable when an AI system continues to learn and evolve, for example. If the future is different from the past, calculations may become problematic. The solution employed by the proponents of the extreme realist position is often to produce even more objective knowledge, albeit acknowledging that it can never be complete.[46] In the context of technical standards, the ISO has attempted to soften this extreme realist perspective by introducing revised terminology where risk is now defined as 'the effect of uncertainty on objectives'.[47]

In the current ISO concept of risk, probability has shifted to likelihood, which is a considerably more open concept; *likelihood* is defined as the chance of something happening, and may be described semi-qualitatively or quantitatively. However, the most important change is the explicit acknowledgement of uncertainty.

The increased focus on uncertainty is commensurate with parts of the more recent risk management literature, which attempts to leave the extreme realist position and recognise doubts and ambiguity.[48] For example, according to Renn, the definition of risk contains the following three elements:[49]

44 ISO, *supra* note 18.

45 Jens O. Zinn, *Introduction: The Contribution of Sociology to the Discourse of Risk and Uncertainty* in Social Theories of Risk and Uncertainty: An Introduction 5 (Jens O. Zinn ed., 2008).

46 *Id.*

47 According to the ISO, uncertainty refers to a 'state, even partial, of deficiency of information related to or understanding or knowledge of an *event*, its consequence or likelihood'. The word 'event' means, according to the ISO, the 'occurrence or change of a particular set of circumstances.' For both definitions, see ISO, *supra* note 18, s 1.1.

48 See, e.g., Terje Aven, Risikostyring: grunnleggende prinsipper og ideer 40 (2007).

49 Renn, *supra* note 42 at 1.

- Outcomes, which have an impact upon what humans value;
- The possibility of occurrence (uncertainty); and
- A formula to combine both prior elements to facilitate prioritisation and decision-making.

The latter understanding of risk focusses less on events and more on outcomes that affect the assets we value. Among the challenges with AI risk is that its impact can emerge in a variety of ways, affecting not only such values as health and safety, but also non-discrimination[50] and other fundamental rights. In light of this broad nature of AI risk, the Proposal combines two risk-based regulatory threads in EU law. On the one hand, it is inspired by safety legislation regulating certain products under the so-called new legislative framework (NLF).[51] The NLF imposes conditions for placing a wide range of products on the EU market, with an emphasis on safety. For example, when using medical devices or toys, the health or safety of persons is crucial. These products can be equipped with AI—creating smart cyber-physical systems, such as care robots or advanced toys—which may generate new risks for health or safety.[52] On the other hand, a second inspiration for the Proposal is arguably the GDPR, especially the Impact Assessment for data protection. The risk-based approach of the Proposal includes consideration of risks for fundamental rights, such as data privacy and non-discrimination. Risks to rights are at stake for example when bias in machine learning can lead to discrimination, as when AI is used to assess applicants during recruitment.[53] By employing the 'risks to rights'

50 Joy Buolamwini and Timnit Gebru, *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*, 81 Proceedings of Machine Learning Research 77-91 (2018); Alexander Tischbirek, *Artificial Intelligence and Discrimination: Discriminating Against Discriminatory Systems* in Regulating Artificial Intelligence (T. Wischmeyer & T. Rademacher eds., 2019).

51 *The 'Blue Guide' on the Implementation of EU Products Rules,* (2016/C 272/01), European Commission (2016).

52 Expert Group on Liability (New Technologies Formation), *Liability for Artificial Intelligence and Other Emerging Digital Technologies* (2019); Eduard Fosch-Villaronga and Tobias Mahler, *Cybersecurity, Safety and Robots: Strengthening the Link between Cybersecurity and Safety in the Context of Care Robots*, 41 Computer Law and Security Review 105528 (2021).

53 Buolamwini and Gebru, *supra* note 50; Tischbirek, *supra* note 50.

approach known from the GDPR,[54] the Proposal is not limited to the technical perspective on risk typical of engineering risk management. Instead, this approach integrates legal assessments and risk assessments as well.[55]

Since diverse assets need to be considered, the combination of the two lines of risk thinking (i.e., the NLF's technical risk perspective and the 'risks to rights' approach) is not necessarily easy. From the technical perspective, assets can include the health of those interacting with a system or a machine. By comparison, in the 'risks to rights' perspective, we are ultimately protecting rights (of persons), which is less objective and introduces norm-based reasoning. As opposed to bodily harm, the violation of rights is not an empirical matter, but a legal one.

Risk assessments typically include the estimation of risk—that is, the assessment of a risk level, such as high, medium or low. The simplest approach for estimating risk combines the likelihood and impact of outcomes in a risk matrix. To do this, the risk manager needs to understand what would count as a high impact, which is both context-dependent and varies depending on the underlying asset at stake. In the technical perspective on risk, outcomes can often be quantified, for example with reference to loss of life or other damages. By comparison, risks to rights are arguably more challenging to quantify and require a more qualitative approach. By nature, a human rights impact is difficult to calculate, and the likelihood of its occurrence depends on many legal questions, rather than on calculations. Thus, the Proposal's notion of risk goes beyond the traditional technical understanding of risk and opens for legal considerations, such as whether and how fundamental rights are affected.[56] For example, parts of the legal framework could ensure that human rights are not infringed, despite some interference. This

54 Niels van Dijk, Raphaël Gellert and Kjetil Rommetveit, *A Risk to a Right? Beyond Data Protection Risk Assessments*, 32 Computer Law and Security Review 286, 289 (2016); Yeung and Bygrave, *supra* note 32.

55 The combination of risk and rights perspectives predates the GDPR; *see,* Thérèse Murphy and Noel Whitty, *Is Human Rights Prepared? Risk, Rights and Public Health Emergencies*, 17 Med. L. Rev. 219 (2009).

56 In Yeung and Bygrave, *supra* note 32 at 10, the authors argue that 'the risk-based approach required by the GDPR necessitates that the data controller undertake a contextual "fundamental rights risk assessment"'.

implies some limits on the degree to which risks can be assessed quantitatively, as discussed in the next section.

5 Risk management rigour

In theory, the use of a rigorous methodology for risk analysis could be a strength of the Proposal compared with legislation that uses no specific methodology. It would certainly be interesting if law-making processes relied on meticulously calculating risks rather than engaging in a political struggle where rights and interests are balanced. However, if we look closer at how the risk-based approach of the Proposal was developed and realised, it is not certain that risks were actually calculated as claimed. Moreover, it is not even clear how many risk levels are included in the Proposal. As regards some of the risk levels, the risk-based approach does not seem to have informed the law-making process at all; it looks like a narrative that was added last minute, almost as an afterthought, as discussed below.

Risk management appears to be a relatively rigorous approach to law-making, at least compared with other law-making approaches that do not follow a specific methodology. This impression of rigour is perhaps most clearly justified by the criteria that distinguish between high-risk and other AI systems.[57] These criteria fit into the first step in risk management. According to ISO 31000, risk management commences by defining criteria for how risk will be measured in terms of consequences and likelihood.[58] The risk criteria are then used in the subsequent risk assessment to rank risks based on risk levels (e.g., high, medium and low risk). It is often challenging to determine clear criteria for how consequences and likelihoods are defined and measured. For example, in the context of the *Titanic*, a sinking of the ship could count as a very severe consequence. By comparison, it is less clear what ought to count as a severe consequence in the broad and ambiguous context of AI risk in Europe.

57 Note that these criteria apply only for AI systems not covered by sectoral product safety legislation, cf. Proposal, *supra* note 1, Art. 6 (1) (a).

58 *ISO 31000:2018: Risk Management – Guidelines*, s. 6.3.4, ISO (2018), available at https://www.iso.org/obp/ui/#iso:std:iso:31000:ed-2:v1:en (last accessed 12 September 2021).

The Explanatory Memorandum to the AIA Proposal explains that under the risk-based method, the Proposal identifies risks 'based on a sector-by-sector and case-by-case approach'.[59] The methodology and criteria for selecting areas and use cases of standalone high-risk AI systems[60] are further explained in the Impact Assessment. To summarise, the Commission went through lists of AI systems that had been flagged as problematic in various reports.[61] They then assessed the 'probability and severity of the harms' to 'determine if the AI system generates a high-risk to the health and safety and the fundamental rights and freedom of persons'.[62]

The criteria[63] for assessment are relatively elaborate and qualitative compared with the rather simple focus on quantifying probability and consequences sometimes found in risk assessments. Overall, the criteria appear relevant, but they are arguably too complex for a purely quantitative approach for calculating risk. Nevertheless, the Explanatory Memorandum claims that risks are 'calculated taking into account the impact on rights and safety'.[64] However, the calculations are nowhere to be found in the published material. Risk management methodology does not necessarily require calculations; risk levels can also be estimated based on a softer, more qualitative approach, as outlined in the Proposal's criteria.[65] Still, if risks were not calculated qualitatively, the claim of risk calculation appears

59 Proposal, *supra* note 1, Explanatory Memorandum to the Proposal 8.

60 *Id.*, Annex III.

61 *Id.*, Annex to the Impact Assessment 41. The criteria for selecting high-risk AI systems listed in Annex III to the Proposal are fairly similar to the criteria that will be used for future changes of that list; see Art. 7 (2).

62 *Id.*

63 *Id.*, Annex to the Impact Assessment, footnote 40. They consider such questions as whether an 'AI system has been used or is about to be used'; 'the extent to which it has caused specified types of harms'; 'the potential of the AI system to impact a plurality of persons'; 'whether potentially adversely impacted persons are dependent on the outcome produced by an AI system' or can opt out; the vulnerability of potentially impacted persons; the reversibility of the outcome produced by an AI system; the availability and effectiveness of legal remedies; 'the extent to which existing Union legislation is able to prevent or substantially minimize the risks potentially produced by an AI system'.

64 *Id.*, Explanatory Memorandum to the AIA Proposal 8.

65 *Supra* note 53. Similar criteria are also to be used for future revisions of the high-risk AI list, see Proposal, *supra* note 1, Art. 7.

to be an exaggeration of the methodological rigour employed in drafting the Proposal. Most likely, the law-making process also considered and sought to balance political, economic, social and other considerations, which are difficult to calculate. Therefore, the lack of risk calculations is not necessarily a flaw. However, the assessment criteria are not formulated to generally estimate the level of AI risk for the whole Proposal, but only to distinguish between high-risk AI and other AI categories. This means that the criteria's focus is on a dichotomy (high risk or not) rather than scale of various risk levels. This dichotomy must have been prevalent for much of the drafting of the Proposal, whereas remaining risk levels seem to have been added later, without similarly elaborate criteria. This can probably be partly explained based on the AIA's legislative history. Initially, the 2020 AI White Paper distinguished only between 'high-risk' and other AI applications,[66] similar to a dichotomy found in Article 35 GDPR.[67] In the 2021 Proposal, there are more risk levels, and the risk-based approach has evolved.

Despite their appearance, the risk levels are not consistently named in the various documents published in conjunction with the Proposal. The Explanatory Memorandum distinguishes three levels of risk, but a press release and the Commission's oral presentation of the Proposal mentions four. According to the Explanatory Memorandum, the Proposal differentiates 'between uses of AI that create (i) an unacceptable risk, (ii) a high risk, and (iii) low or minimal risk'.[68] The press release agrees on the two highest levels but splits (iii) into 'limited' and 'minimal risk'. Limited risk is the category of AI systems that includes chatbots, which are subject to transparency obligations (Proposal, Article 52), and minimal risk covers all remaining AI systems.[69] Thus, it is unclear whether there are three or four risk levels.

This is a minor inconsistency, but given the focus on the risk-based approach in the public presentation of the Proposal, it is surprising that the risk levels are not enumerated consistently and included in the Proposal, which could have been done in the recit-

66 *White Paper on Artificial Intelligence*, *supra* note 8 at 17.

67 GDPR, *supra* note 16, Art. 35 (1).

68 Proposal, *supra* note 1, Explanatory Memorandum accompanying the AI Proposal 12.

69 *EU Fact Sheet*, *supra* note 14.

als. The multiple levels of risk in the Proposal seem inspired by a 2020 report[70] from the German Data Ethics Commission, which developed a well-defined 'criticality pyramid' with five risk levels of algorithmic systems. In the pyramid, level 5 criticality indicates algorithmic applications with 'untenable potential for harm', which should be banned, whereas level 1 denotes applications with zero or negligible potential for harm, requiring no special measures. By comparison, the 2021 AIA Proposal reads as if it were conceptually based on a dichotomy between high-risk and other AI (as in the White Paper). As mentioned above, further risk levels seem to have been added as an afterthought when additional categories of rules had emerged in the law-making process, and not necessarily informed by the same risk criteria.

This view is strengthened by the fact that the prohibited AI practices do not carry an explicit risk level in the Proposal. Instead, Article 5 simply enumerates prohibited AI practices. In addition, the definitional scope of the various rules differs. Whereas Article 5 prohibits certain AI *practices*, Article 6 classifies certain AI *systems* as high risk (and thus regulated). This means that the various rule sets do not differ only in terms of risk criticality, but also in substantive scope.

The Proposal includes relatively elaborate criteria for distinguishing between high-risk AI and other AI, but there are no explicit criteria available for the remaining risk classes. On the other hand, the aforementioned criteria would arguably be sufficiently general to be used for assessing all risk levels, so they could have been applied more broadly, as well as for discussing the levels of 'unacceptable risk' and 'limited risk'.[71]

The criteria are also relevant for discussing the coherence of risk assessments in the Proposal. One example is the risk level assigned to two different AI systems, both related to the creation of deep fakes, which result in different risk levels. First, AI systems that generate deep fakes are in the 'limited risk' category (although no such category is explicitly mentioned in the Proposal). They are only

70 Data Ethics Commission (Germany), *Opinion* (2020), available at http://www.odbms.org/2020/10/opinion-of-the-german-data-ethics-commission/ (last accessed 12 September 2021).

71 Potentially, some of the criteria could also be relevant for risk assessments under Art. 9.

subject to transparency obligations.[72] This classification in the second-to-lowest risk category is noteworthy because a similar type of AI system is in a different class. 'AI systems intended to be used by law enforcement authorities to detect deep fakes' are high risk—that is, one risk level up.[73] In other words, deep fakes are considered less risky than the tools used to detect them, which is surprising. Explicit risk criteria could be used to inform and evaluate these risk analyses, and overall, the criteria included in the Proposal are a good starting point.

6 Conclusion: Risk and legislative proportionality

By way of conclusion, let us return to some of the questions mentioned in the introduction, focussing on the risk-based approach's function, relevance and limitations. The European Commission emphasises that it 'is proposing new rules to make sure that AI systems used in the EU are safe, transparent, ethical, unbiased and under human control. *Therefore,* they are categorised by risk' (emphasis added).[74] In other words, this indicates that the risk-based approach is used mainly to manage risks. However, the Commission also argues that the main alternative to this risk-based approach would have been the 'blanket regulation of all AI systems'.[75] Such an alternative would not have implied that the risks of AI systems are ignored; AI risk would surely remain a regulatory rationale, but AI innovation might be hampered. An overly extensive regulation of AI might disproportionately hamper the development of societally desired and economically lucrative AI.[76]

The above suggests that the risk-based approach is not, in fact, primarily an attempt to manage risks, but a solution to a specific challenge in the Proposal—namely, the very broad definition of AI

72 AI Regulation, *supra* note 1, Art. 52(3).

73 *Id.*, Annex III, 6(c).

74 *EU Fact Sheet*, *supra* note 14.

75 *Ibid*. The Commission has also considered other alternatives, as described in the Impact Assessment.

76 *Innovative and Trustworthy AI: Two Sides of the Same Coin*, Position Paper on Behalf of Denmark, Belgium, the Czech Republic, Finland, France, Estonia, Ireland, Latvia, Luxembourg, the Netherlands, Poland, Portugal, Spain and Sweden.

systems. Already in the White Paper, the risk-based approach was considered important to help ensure that the planned regulatory intervention would be proportionate.[77] Hence, high-risk AI applications were subject to requirements, whereas below-threshold AI applications were largely exempted from the additional requirements envisaged in the White Paper. The risk-based approach seems to have primarily emerged as a mechanism to exclude certain types of unproblematic AI systems from the AIA's scope.

Defining AI is a classic problem, partly caused by the challenge that we do not even have a good understanding of natural intelligence. In the Proposal, the term *AI system* 'means software that is developed with one or more of the techniques and approaches listed in Annex I and can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations or decisions influencing the environments they interact with'. This encompasses a wide range of software,[78] ranging from spam filters to algorithms used in lethal autonomous weapons, which raise vastly different regulatory issues. Therefore, regulating all AI systems would render an excessively wide scope of the Regulation, even considering any limitations that may follow from Annex I. The primary function of the risk-based approach is simply to limit the scope of a potentially over-broad regulation by tailoring the rules to the 'intensity and scope of the risks that AI systems can generate'.[79]

This raises the question of whether the risk-based approach might also be available and relevant for other policy domains. In my view, this option exists, but it is not necessarily clear whether the approach has a significant effect on the regulatory framework in which it is applied. This can be illustrated by hypothetically inserting the risk-based approach into a different ruleset, such as rules regulating the use of land-based motorised vehicles. Existing rules in this area could be kept unchanged; the risk-based approach would only apply a new framing.

A hypothetical risk-based approach to vehicles could be structured as follows:

77 European Commission, *supra* note 8 at 17.

78 Michael Veale and Frederick Zuiderveen Borgesius, *Demystifying the Draft EU Artificial Intelligence Act*, 22 Computer L. Rev. Int'l. 97 (2021).

79 AI Regulation, *supra* note 1, Recital 14.

- *Unacceptable risk:* Vehicles representing an unacceptable risk are prohibited. This applies, for example, to tanks, except when used by military forces.
- *High-risk vehicles*: The use of vehicles with engine size exceeding a certain threshold (cars, trucks, etc.) is regulated.
- *Limited risk:* Specific rules apply to certain categories of vehicles, such as electric scooters and mowers.

This use of a risk-based approach is possible, but it would not necessarily be considered an improvement of existing vehicle rules; it would essentially be limited to a re-branding of rules. Similarly, the Proposal could have been drafted (and branded) without the risk-based approach by simply creating a category of 'regulated AI systems' (perhaps with a better name), corresponding to the current use of the term 'high-risk AI systems'. This might also have avoided some potential confusion about AI risk levels in the concrete application of the law, where 'high-risk AI systems' can paradoxically involve 'low risk'.[80] Whereas the risk-based approach may be suitable for achieving proportionality and avoiding regulatory overreach, this comes with a potential cost, as it could cause confusion. Risk levels need to be assessed at various levels and by various actors, including the law-maker, AI providers and national authorities. Therefore, risk levels are essentially context-dependent and are not necessarily consistent across the various risk assessments. They depend on the risk criteria of the actor carrying out the respective risk assessment. A *type of* AI system may be considered 'high-risk AI' from a regulatory perspective, but this does not mean that the *concrete* AI system involves 'high risks' when these are assessed by the AI provider or an authority.

It might have been easier to reserve risk levels primarily for the risk analysis of concrete AI systems while using a different nomenclature for distinguishing prohibited, highly regulated and less regulated AI systems. In contrast, the explicit criteria for high-risk AI systems are an interesting innovation that could become a model for other future law-making processes. Moreover, the criteria could have been applied more consistently across all risk levels addressed in the Proposal, which would have strengthened the logic of the overall law-making process intended to produce the final EU AIA.

80 See above, Section 2.

Part 4

How to Regulate?

Can Artificial Intelligence be Regulated? Lessons from Legislative Techniques

UBENA JOHN

Abstract
Artificial intelligence (AI) has become an integral part of human life. It has led to the emergence of new services, goods, and apps. The latter have helped to address some societal challenges in various sectors, including the health sector. eHealth has been adopted to overcome a shortage of medical experts and improve health services, monitor pandemics, and increase mobility. On the other hand, unregulated AI, such as lethal autonomous weapon systems, creates threats of uncontrollable arms. Moreover, AI systems are difficult to regulate because the legislature and regulators have not grasped their developments and uses. While a haphazard regulatory intervention may stifle AI innovation, the dangers posed by AI are to some extent known. Several uncoordinated initiatives at a national or regional level have been deployed to investigate legal, ethical, and other regulatory challenges of AI. A Comprehensive global legal framework to deal with AI issues is lacking, except for the proposed EU AI Regulation (2021). This work argues for the adoption of legislative techniques in the regulation of AI. These techniques allow oversight of AI development and serve to define rights, obligations, liabilities, and remedies. Since the techniques operate as a regulatory toolbox, a regulator may pick an appropriate technique to address a particular regulatory challenge related to AI. There is no one size fits all approach. The techniques complement each other, as opposed to operating in isolation.

1 Introduction

This chapter contributes to the discourse on the regulation of artificial intelligence (AI). It examines how AI may be regulated using legislative techniques. Recent debates have shown that AI has created digital personal assistants, translation aids, facial recognition systems, and other expert systems to support commerce, public administration, and health care provision. In the latter case, this includes medical records, diagnosis, treatment and diseases surveillance, prevention, and control systems, among other things. Robotics is another field of AI that is growing, having brought driverless cars, unmanned aircrafts, and others to the market.[1] While AI is beneficial, it suffers a lack of transparency and clear legal and regulatory frameworks. Thus, it creates ethical and legal challenges, including threats to certain legal rights and state relations. Use of AI systems can mean that responsibility and liability for errors, for instance in autonomous cars, lethal autonomous weapon systems (LAWS), are difficult to attribute to a certain entity. To further attempts to regulate AI, the chapter briefly analyses the legislative techniques (LTs) and how they may be applied in regulation of AI systems. It is submitted that AI can be regulated if regulators make use of the LTs, because they encompass various regulatory tools and mechanisms that are multimodal and flexible enough to match AI's complex and unpredictable trends.

This chapter begins with definitions of AI and LTs as extracted from the literature. It then proceeds to outline the potentialities of and legal issues related to AI. The next section attempts to analyze the application of LTs in the regulation of AI. Lastly, concluding remarks and areas for future investigation are presented.

1.1 What is AI?

Artificial intelligence can mean the use or creation of intelligent machines or systems. In the Ethics Guidelines for Trustworthy Artificial Intelligence, the High-Level Expert Group on Artificial Intelligence (AIHLEG) defines AI as "systems as software (and possibly also hardware) systems designed by humans that, given a complex goal, act in the physical or digital dimension by perceiving their

1 Ralf T. Kreutzer & Marie Sirrenberg, Understanding Artificial Intelligence: Fundamentals, Use Cases and Methods for a Corporate AI Journey at 1 (2020).

environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge, or processing the information, derived from this data and deciding the best action(s) to take to achieve the given goal."[2] The field includes machine learning, machine reasoning – such as planning, scheduling, and knowledge representation – and robotics, which includes control, perception, and sensors and actuators. Robotics may also involve integrating AI into cyber-physical systems.[3]

1.2 What are Legislative Techniques?

The term 'legislative technique' is used here in its generic sense, to encompass legislative drafting techniques, mechanisms for implementing and enforcing law, and the New Regulatory Culture. The latter encompasses self-regulation, co-regulation, involvement of non-state actors in regulation, the use of code is law, contracts, and transformation of prescriptive behavior norms into duty of care norms.[4]

Why legislative techniques? This is a good question. It should be mentioned that regulation tends to control the behavior of the regulated entities. Moreover, regulation restricts certain freedoms and

2 High-Level Expert Group on Artificial Intelligence (AIHLEG), *Ethics Guidelines for trustworthy AI*, European Commission, (2018) at 36; Communication from the European Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of Regions, *Artificial Intelligence for Europe*, COM/2018/237.

3 High-Level Expert Group on Artificial Intelligence (AIHLEG), *A Definition of AI: Main Capabilities and Scientific Disciplines*, European Commission (December 18, 2019), https://ec.europa.eu/futurium/en/system/files/ged/ai_hleg_definition_of_ai_18_december_1.pdf (last accessed May 30, 2021).

4 Ubena John, How to Regulate Information and Communications Technologies? A Jurisprudential Inquiry Into Legislative and Regulatory Techniques 4–8 (2015). *See also* Peter Wahlgren, Lagstiftning: Rationalitet, teknik, möjligheter (2014); Peter Wahlgren, Lagstiftning: Problem, teknik, möjligheter *(i.e., Legislation: Problem, Techniques, Possibilities)* (2008); Peter Wahlgren, *IT and Legislative Development*, 47 Scandinavian Studies in Law 601–617 (2004); *see also* Willem van der Velden, *The Value-and Goal-Dependency of Legislation and Its Methodology*, in The Structure of Law 53–100 (Åke Frändberg & Mark Van Hoecke (eds.), 1987); Constantin Stefanou & Helen Xanthaki, Drafting Legislation: A Modern Approach (2008); Helen Xanthaki, Thornton's Legislative Drafting (2013); Helen Xanthaki, Legislative Drafting: Art and Technology of Rules for Regulation (2014).

thus may act as a constraint on the regulated entities. Restriction of rights and freedoms must be lawful, legitimate, necessary, and proportionate. From the perspective of democratic states, the legislature has been given a monopoly to make laws. Legislative techniques are tools for legislature and regulators. Since the latter bodies are mandated to make laws, once they adopt the LTs, the questions of lack of legitimacy, lawfulness, and proportionality will rarely arise.[5]

1.3 What are AI potentialities?

As already mentioned, AI has many potentialities. In Australia, for example, an AI system has been developed to support tree planting. This is one way to address pressing global climate change through afforestation. Climate action and protection of natural resources are among the UN Sustainable Development Goals.

AI has also contributed to improved healthcare service provision. Thus, AI enhances universal health coverage via AI-based eHealth systems. The applications include AI for diagnostics, real-time monitoring of a patient's condition, treatment, and disease control. Technologies based on AI and robotics can be valuable tools or assist caregivers in pandemics of COVID-19 or Ebola, for example.

Moreover, AI in the field of the robotics has proved to be beneficial for society. For instance, robots may be deployed as first responders in emergency situations like natural disasters. Drones have played a vital role in transportation of paramedics during floods and similar emergencies. Robotics has further contributed to the development of autonomous vehicles. Autonomous cars have improved mobility, for instance for handicapped individuals.

1.4 What are the legal issues?

Despite the potentialities, AI creates several ethical and legal challenges. Complicating this further is the fact that the legislature and regulators have not agreed upon how an AI legal and regulatory framework should be constructed. The legal challenges include new AI apps and technologies that have made certain laws obsolete (civil

5 Ubena, *supra* note 4 at 109–121; Robert S. Summers, *The Technique Element in Law*, 59 Calif. L. Rev. 733-751 (1971); R.F. Cranston, *Reform through Legislation: The Dimension of Legislative Technique*, 73 Nw. U. L. Rev. 873, 873–908 (1979); Luc Wintgens, The Theory and Practice of Legislation: Essays in Legisprudence (2005).

aviation laws and road traffic laws cannot apply to drones and autonomous vehicles without modification); affected consumer interests (e.g., reimbursement for AI-based medical services, including treatment); unfair competition (such as Uber services that threaten the traditional taxi business); and goods being transformed into services and services into goods (Platform as a Service, Software as a Service, Infrastructure as a Service, etc.). This last transformation blurs the distinction between goods and services, which makes regulation difficult, especially in case of technology- or service-specific regulation.

It is generally understood that traditional law is penal – based on command and control – and coercive by nature.[6] Although traditional regulations and laws were intended to control the behavior of individuals, they were based on definitions, technology, and a service divide.[7] The regulations were also imposed without an understanding of the regulatory environment, which risked stifling innovation. The assumption of traditional law was that the wrongdoers were known or would be identified and could be punished. Offences or wrongs were known or foreseen before their occurrence. But what about intelligent machines, robots or other AI apps that are autonomous or semi-autonomous and may commit or cause offences to be committed? Can they be regulated by traditional laws?

In addition to the above controversy, operation of an AI often lacks transparency.[8] That may be due to the complexity of the underlying algorithms. One might ask whether an algorithm's basic data should be published.[9] Consider for example a contract concluded with the help of an algorithm, i.e., an algorithmic contract (based on embedded standard terms). While similar to a standard form contract, such a contract may lead to questions on transparency and

6 Hans Kelsen, *Law As a Specific Social Technique*, 9 U. Chi. L. Rev. 75, 75–97 (1941).

7 Ubena, *supra* note 4 at 43–47, 181–233.

8 Wolfgang Hoffmann-Riem, *Artificial Intelligence As a Challenge for Law and Regulation*, in Regulating Artificial Intelligence at 17 (Thomas Wischmeyer & Tim Rademacher (eds.), 2020); Thomas Wischmeyer, *Artificial Intelligence and Transparency: Opening the Black Box*, in Regulating Artificial Intelligence (Thomas Wischmeyer & Tim Rademacher (eds.), 2020).

9 *See* Tanel Kerikmäe & Katrin Nyman Metcalf, *Machines Are Taking Over – Are We Ready?* 33 SAcLJ 24, 39-41 (2021).

consumer protection, if concluded without the makeup of the algorithm being disclosed.[10]

Moreover, AI may fail to comply with certain existing laws. For instance, the GDPR requires that data processing principles be respected. One of the principles is that the purpose of collecting and processing personal data should be specific, explicit, and legitimate.[11] This may be difficult to adhere to in AI systems, as their operations are dependent upon processing huge quantities of data.

Apart from that, there are also intriguing questions surrounding AI: will or can AI respect Asimov's laws of robotics? What happens when AI disrespects these laws?[12] Asimov's laws of robotics say:

> 1st A robot may not injure a human being; 2nd A robot must obey the orders given to it by human being except where such orders would conflict with the First Law; 3rd A robot must protect its own existence as long as such protection does not conflict with the First and the Second Laws; and 4th A robot may not injure humanity, or by inaction allow humanity to come to harm.[13]

Though Asimov's laws of robotics or Haley's metalaw may be relevant in AI regulation, in the strict sense, this chapter deals with neither of them. Rather, it deals with law and LTs as applied in the regulation of AI. And, as rightly stated by Pagallo: even angels need the rules.[14] This statement implies that no matter how perfect AI may be, legal rules for regulating their conduct are inevitable. One could argue that these rules are intended to serve humanity. However, autonomous AI systems – if not controlled – may pose threats. Thus, it is unsurprising that AIHLEG, in their study of ethical and

10 *Id.* at 41.

11 European Union, Council Regulation 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC, OJ 2016 L 119/1 [*hereinafter* GDPR], Article 5(1)(b).

12 *See,* in this book, a chapter by Chris Reed (2021); Kamil Muzyka, *The Basic Rules for Coexistence: The Possible Applicability of Metalaw for Human-AGI Relations*, 11 Paladyn, J. Behav. Robotics 104-117 (2020). Both have examined Asimov's laws of robotics.

13 *See* Muzyka, *id.* at 105.

14 Ugo Pagallo, *Even Angels Need the Rules: AI, Roboethics and the Law*, available at https://ebooks.iospress.nl/volumearticle/44760 (last accessed May 30, 2021).

legal AI, observed that in most AI systems that have been developed and used, there is lack of human-centric AI.[15]

The need for human-centric AI controlled by legal rules stems from a confusion regarding how the rights and liability of AI players have been defined, and the fact that AI is unpredictable.[16] For example, who should bear responsibility for errors or injuries caused by an AI-based diagnostic and treatment system? Can medical malpractice and product liability rules apply in such a situation?[17] Such human-centric ethical and legal issues have not been reflected in AI systems.

Moreover, AI systems create ambiguity in apportioning responsibility, e.g., who is at fault – who is to be blamed in case of a road accident involving autonomous cars? Who is responsible for an incorrect diagnosis caused by a dysfunctional AI-based mHealth or eHealth monitoring or diagnosis system? How is patient autonomy or informed consent to be realized in AI-based health services?[18] It may be interesting to examine how AI-based systems for health services have changed the legal relationship between physicians and patients.[19] Can or should ethics of professional conduct or the doctor-patient contractual relationship (duty of care during treatment) apply to AI systems deployed in the provision of health care? Do medical liability, contractual liability, and tortious liability apply to AI systems in healthcare? The lack of information provided to patients (affecting patients' informed consent) and treatment errors may lead to application of medical malpractice law. But does this apply to AI-based healthcare? Who should be responsible if there is an error in an AI-based medical device? The physician or the device

15 AIHLEG, *supra* note 2.

16 Daniel Schönberger, *Artificial Intelligence in Healthcare: A Critical Analysis of the Legal and Ethical Implications*, 27 Int'l. J.L. & Info. Tech. 171–203 (2019) (advocating for the use existing legal frameworks, e.g., product liability, negligence, medical law, etc., to deal with AI issues in eHealth services).

17 Some of these questions have been explored by Liane Colonna, *Lifelogging Technologies For the Frail and Sick*, 27 Int'l. J.L. & Info. Tech. 50–74 (2019) (lifelogging technologies in the context of data protection).

18 Fruzsina Molnár-Gábor, *Artificial Intelligence in Healthcare: Doctors, Patients and Liabilities*, in Regulating Artificial Intelligence at 338–358 (Thomas Wischmeyer & Tim Rademacher (eds.), 2020).

19 Some scholars have already done so. *See* Schönberger, *supra* note 16.

manufacturer?[20] In this context, product liability questions may arise. Interestingly, on the responsibility gap, Jan-Erik Schirmer proposes granting legal personality to AI systems (intelligent agents). According to him, this would fill or close the responsibility gap.[21] However, doing so may have other, unintended consequences. For example, it may cause people to question why these intelligent agents should not be granted the equal rights that natural persons enjoy. Moreover, should the intelligent machines be held criminally responsible? What penalty should be imposed upon them?

Traditionally, law grants rights, imposes obligations or duties and defines offences. Thus, granting of AI-related rights falls under the *traditional legislative techniques*. In keeping with this, as a remedy for the errors caused by intelligent agents, they could be reprogrammed or be shut down.[22] Kerikmae and Metcalf have also discussed the possibility of law granting legal personality to AI systems.[23] Although there have been debates on granting AI systems legal personality, there has also been a fear of robots seizing power over human beings or of robots not obeying humans.[24] Still, the granting of legal personality to inanimate things is not new. The legislative assembly of El Salvador in 2019 granted rights to forests as living entities.[25] New Zealand has granted legal personality to the river Whanganui.[26] Nonetheless, granting AI legal personality may give rise to a question as to its rights and duties. Moreover, is AI capable of acting consciously?[27]

20 For more on AI and Healthcare, *see* Sarah Jabri, *Artificial Intelligence and Healthcare: Products and Procedures*, in Regulating Artificial Intelligence 307–335 (Thomas Wischmeyer & Tim Rademacher (eds.), 2020).

21 Jan-Erik Schirmer, *Artificial Intelligence and Legal Personality: Introducing "Teilrechtsfähigkeit": A Partial Legal Status Made in Germany*, in Regulating Artificial Intelligence 124–141 (Thomas Wischmeyer & Tim Rademacher (eds.), 2020).

22 *Id.* at 139.

23 *See* Kerikmae and Metcalf, *supra* note 9 at 43–46.

24 *See* Max Tegmark, Life 3.0: Being Human in the Age of Artificial Intelligence (2017).

25 *El Salvador recognizes forests as living entities*, Down to Earth (June 11, 2019), https://www.downtoearth.org.in/news/forests/el-salvador-recognises-forests-as-living-entities-65020 (last accessed May 30, 2021).

26 *See* Kerikmae and Metcalf, *supra* note 9 at 44–45.

27 *Id.* at 45.

Regardless of the legal challenges, AI apps have been used for identifying and tracking people for various reasons. They have been used to track elderly people, patients with mental diseases, and even those suspected of having a COVID-19 infection. This gives rise to a question of Big Brother surveillance and related privacy violations. However, the legal framework for this scenario seems to exist within the EU. There is the EU GDPR, which is technology-neutral. Hence, it could apply to AI.[28] On the other hand, covert or "secret" AI systems have attracted the attention of governments. Developers of AI systems should ensure that humans are informed when they are interacting with AI systems, rather than with other humans. This might become a serious concern, if the development of human-like robots proceeds. From a standpoint of ethics and fundamental rights, humans should be given the opportunity to decide whether or not to interact with a robot[29]. AIHLEG has also raised the alarm regarding AI-based scoring systems, such as automated grading in education/schooling systems or deducting points on a driver's license (speed cameras); these should be used carefully and only in line with fundamental rights.[30]

Lethal autonomous weapons systems (LAWS) are another source of controversy. They are finding their way into the weapons industry and some countries have invested in and developed such weapons and missiles. These systems can make decisions and operate without human intervention. We are on the verge of entering into an arms race that can hardly be controlled by human beings.[31] Should disarmament apply to these LAWS? How much do we know about them? Should there be an international convention to outlaw such weapons? Application of international laws, international humanitarian laws, and oversight and control of LAWS are essential.

In addition to LAWS, unmanned aircrafts, while especially useful in emergency situations, can be used to transport weapons or to attack someone. A good example that some may recall is the drone attack on President Maduro. This was not the fault of the drone,

28 *See* AIHLEG, *supra* note 2 at 33–34.

29 *Id.* at 34.

30 *Id.*

31 *Id.*

which is a neutral device.[32] Rather, it was the fault of a person who used it to transport a bomb aimed at Maduro. As this example shows, AI-based drones – if unregulated – may fall into the wrong hands or be used for criminal activities. To resolve that problem, regulators could impose licensing requirements on drone users, as shown elsewhere in this chapter.

2 What is the role of Legislative Techniques in the regulation of AI?

We have seen how problematic AI and its regulation can be. This section turns to the next question: how can LTs be applied in the regulation of AI? Legislative techniques entail legislative-centric, traditional command and control measures (using the coercive apparatus, e.g., the police and the prisons), criminalization, penal sanctions, lawmaking, legislative drafting, institutions, and tools/mechanisms for enforcing legislation. They also encompass other regulatory techniques, e.g., code is law, i.e., computer programs embedded into the law for effective enforcement.[33]

The *traditional legislative techniques* (TLTs) represent traditional law (command and control) – control of behavior through law, also called Law 1.0. It is the law in traditional sense, i.e., law aimed at commanding and controlling the behavior of the regulated entities. Non-compliance is redressed through penal sanctions. For instance, the application of TLTs in AI regulation will entail crafting a rule that AI must behave in certain way, e.g., to secure the privacy of patients or consumers. If it fails, the algorithm must be reviewed, to identify the error and reprogram the AI system. In other instances, and as a last resort, the AI may be banned or shut down.[34] However,

32 *See* Nick Paton Walsh et al., *Inside the August Plot to Kill Maduro with Drones*, CNN (June 21, 2019), https://edition.cnn.com/2019/03/14/americas/venezuela-drone-maduro-intl/index.html (last accessed May 30, 2021). Maduro is the President of Venezuela since 2013.

33 Ubena John, *The Role of Legislative Techniques in Regulation of Disruptive Technologies*, The Loophole – Journal of the Commonwealth Association of Legislative Counsel 1-21 (2019); Ubena, *supra* note 4 at 4–8; Lawrence Lessig, Code and Other Laws of Cyberspace 85–99 (1999).

34 *See* Kerikmae and Metcalf, *supra* note 9 at 43–46; Schirmer, *supra* note 21 at 128–132, 138–139. *See also* Article 5 of the Regulation of the European Parliament and of the

this will require AI to be granted legal personality. Therefore, these systems will have rights and obligations. This has not been done yet. As shown in other sections of this chapter, the possibility has already been discussed by other scholars. Scholars have also examined the role of AI as an agent of human beings or the institution that uses that AI system.[35]

In addition to employing a coercive apparatus in law enforcement, the TLTs depend on how the legislation is drafted. Therefore, drafting style matters. In the TLTs, the legislative drafting style adopted is either detailed (some AI systems, e.g., eHealth and mHealth apps, require this approach) or non-detailed (other AI systems, e.g., blockchain, Internet of Things appliances, require this approach, because they are always changing).

Detailed drafting takes onboard technology-specific laws, criminalization and banning of certain behavior, but also entails limiting the scope of application of a law through definitions[36], uses and technologies.

While detailed drafting may seem to be narrowly focused and restrictive, it is desirable in certain circumstances, for example, to enhance cybersecurity and to prevent discrimination.[37] One can consider an AI system developed in the USA for automation of judicial decision-making. The system was seen to be discriminatory against Black African American suspects/accused persons (who got

---

Council Laying Down Harmonized Rules on Artificial Intelligence (Artificial Intelligence Act) And Amending Certain Union Legislative Acts (COM (2021) 206 final) hereinafter referred to as the Proposed EU AI Regulation (prohibits certain types of AI practices, e.g., AI systems that exploit any vulnerabilities of a specific group of persons due to age or physical or mental disability).

35 *See* Kerikmae and Metcalf, *id.* at 34, 43–45; Schirmer, *id.* at 125.

36 A good example here is the concept of 'medical device' under the EU Medical Devices Directive (Directive 93/42/EEC). To determine whether a particular device is a medical one, one must consider the underlying purpose and whether its use was prescribed by a physician. Therefore, most wellness devices and apps on the market (including AI-based eHealth apps) may be excluded from the definition of medical device. This poses challenges in determining consumer rights, as well as the obligations and liabilities of eHealth device manufacturers, suppliers and service providers.

37 *See* Hoffmann-Riem, *supra* note 8 at 13.

high ratings) in comparison with White suspects/accused persons (who got low ratings).[38]

2.1 The New Regulatory Culture

Traditional legislative techniques have evolved into the regulatory toolbox known as the New Regulatory Culture (NRC). The term NRC was first used in the EU Better Regulation strategy to mean a mixture of the evolved TLTs and other emerging regulatory strategies in the regulation of a particular thing.[39] The NRC does not supplant TLTs, but rather complements them. The NRC is goal-steering.[40] It is result-oriented. It extends from modernized TLTs to other approaches,[41] such as sunset laws, embedded solutions in a form of code is law, contract-based strategies, proactive approaches based on incentives, regulatory forbearance to evolutionary approaches based on duty of care norms, liability rules, and regulatory impact assessments.[42]

38 Julia Angwin et al., *Machine Bias*, ProPublica (May 23, 2016), www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (last accessed May 30, 2021); Gabriele Buchholtz, *Artificial Intelligence and Legal Tech: Challenges to the Rule of Law*, in Regulating Artificial Intelligence 189–190 (Thomas Wischmeyer & Tim Rademacher (eds.), 2020).

39 *See* the *Report on Implementation of the European Commission's Work Programme for 1996*, European Commission, (October 16, 1996); The Report on the Introduction of Better Regulation in the EU, 'the Mendelkern Group on Better Regulation Final Report' of November 13, 2001. Available at https://www.smartreg.pe/reportes/Mandelkern%20Report%200n%20Better%20Regulation%202001.pdf (last accessed June 7, 2021); Linda Senden, *Soft Law, Self-Regulation and Co-Regulation in European Law: Where Do They Meet?* 9 Electronic J. Comp. Law 1–5 (2005); Koen Van Aeken, *Legal Instrumentalism Revisited*, in The Theory and Practice of Legislation: Essays in Legisprudence 77–88 (Luc J. Wintgens (ed.) 2005); Ubena, supra note 4 at 4–8, 237–238; Pauline Westerman, *Governing by Goals: Governance as a Legal Style*, 1 Legisprudence 51–72 (2007); Pauline Westerman, *Who is Regulating the Self? Self-Regulation as Outsourced Rule-Making*, 4 Legisprudence 225–241 (2010).

40 Pauline Westerman, *Breaking the Circle: Goal-Legislation and the Need for Empirical Research*, 1 Theory & Practice of Legislation 395–414 (2013).

41 In this chapter, the words techniques and approaches have been used interchangeably.

42 *See* Ubena, *supra* note 4 at 78-87; Ubena, *supra* note 33. *See also* Hoffmann-Riem, *supra* note 8 at 18, 22–25 (advocating for the adoption of hybrid regulation in regulating AI).

What indicates that NRC does not supplant TLTs is that an increasing adoption of non-detailed legislative drafting is currently being seen. Such a drafting style provides discretion and flexibility to national regulatory authorities in developing rules to implement legislation, especially where the regulatory environment is new, technical, and constantly changing. Non-detailed legislative drafting allows the use of co-regulation and, in some instances, self-regulation – which are both central to the NRC.

2.1.1 Functional equivalence principle

Non-detailed drafting also supports the functional equivalence principle, in the sense that what applies offline also applies online. It may also mean that what applies to other technologies should apply to AI. One could question if the rules of medical professional negligence should apply to AI for health systems. The functional equivalence principle is not without critics. The principle may not apply where technologies are dissimilar. Further, there are some areas or rights that do not have offline equivalences, e.g., the right to be forgotten or anonymity.[43] In such instance, new rights should be created to suit AI.

2.1.2 Technology-neutral laws and AI

Several scholars have examined and advocated for technology-neutral laws.[44] While they have looked at it from different perspectives, one thing they all agree on is that laws should be neutral, instead of targeting or mentioning a specific technology. That resolves the problem of obsolescence of law when new technologies emerge. A good example is GDPR Article 5(1)(b), which applies across various technologies. However, data protection principles, in particular the data minimization principle, i.e., that data can only be processed for a certain specified and legitimate purpose and must be deleted as soon as they are no longer relevant to the purpose, may restrict

43 *See* Ubena John, *Digital Rights in Africa in the Era of COVID-19 and Beyond*, in 20 Years of Cyberlaws in Africa (Mohamed Chawki and Sizwe Snail (eds.), 2021) (forthcoming).

44 Chris Reed, *Taking Sides on Technology Neutrality*, 4 SCRIPT-ed 263–284 *(* 2007); Bert-Jaap Koops, *Should ICT Regulation be Technology-Neutral*, in Starting Points for ICT Regulation-Deconstructing Prevelant Policy One-Liners 77–108 (Bert-Jaap Koops et al. (eds.), 2006).

AI as it requires large amounts of data to be used efficiently. Therefore, in relation to this principle, the GDPR may be a constraint to AI development.[45] However, if consent for data processing is given, there may not be any problem even when data processing is for more than one purpose, as is the case with AI systems.

Kerikmae and Metcalf have analyzed the regulation of AI, including the use of technology-neutral laws as opposed to technology-specific laws, adapting, or amending the laws to reflect AI application in public administration. This is crucial to support e-Government projects.[46] These scholars went further, to discuss the responsibility for AI. If there is damage caused by an AI decision, the state (government) should be responsible because AI apps are used as agents of the state. From these scholars' discussion, one can note that they excluded non-state (public administration) AI decisions, e.g., commercial transactions that are purely private matters. A question that can be asked here is who should be responsible if an AI-based mHealth or eHealth app's malfunction leads to an incorrect diagnosis, e.g., showing high blood sugar levels when the levels are actually low. In such a situation who should be blamed: the system developer who developed the app or the health care provider requiring or recommending the use of the app? The same applies to driverless cars and other AI-based apps, goods, and services. Scholars have proposed several approaches to deal with these issues. Some have suggested relying on liability rules and the manufacturer's liability.[47] The challenge is the involvement of AI system that may in certain instances be autonomous. Who should be responsible in case of injury caused by such AI systems? It is for this reason that scholars have recommended adoption of LTs combining various tools that could enable the legislature to understand the regulatory environment (i.e., the AI system). Depending on the nature and purpose of the regulation, some approaches allow trial and error, while others require impact assessments and a shift from behavior norms to duty of care norms.

45 *See* Kerikmae and Metcalf, *supra* note 9 at 42. It is also important to consider the changes that may be brought about by the EU AI Regulation.

46 Id.

47 *See* Kerikmae and Metcalf, *supra* note 9 at 42; Hoffmann-Riem, *supra* note 8 at 12.

2.1.3 Techniques for the New Regulatory Culture

It is also possible to use technology to overcome AI legal and regulatory challenges. From the perspective of legislative techniques, this is referred to as the *technological approach*. Other scholars have named it "code is law" or Law 2.0, i.e., embedded solutions.[48] In that approach, other concepts – such as privacy by design – also emerge. Article 25 of GDPR provides for data protection by design. The technological approach has gained recognition from expert groups. For instance, the AIHLEG in the EU has suggested the programming of trustworthy AI – embedding trustworthy AI features into AI, i.e., rule of law by design, privacy by design, and security by design.[49]

Another approach is *complementary*, or contract-based (self-regulation, codes of conduct, certifications, standards, etc.). The parties (regulated entities) are bound by the contractual terms (sanctity and freedom of contract). While contracts were originally matters between private parties, they have now found their way into the legislation. They are referred to as enforced self-regulation.[50] The GDPR has a provision that requires data controllers and processors to have a code of conduct. Similar approaches may be adopted for AI regulation, as proposed in the EU AI Regulation (Article 69), where the AI developers and service providers may have a code of conduct; anyone who violates it may be expelled from the AI community.

The *proactive approach* is also included among the LTs. This approach encompasses economic incentives, regulatory forbearances, nudging, insurance schemes, and social strategies: educational programs, information supply, etc. Ideally, the proactive approach aims at pre-empting legal problems. Therefore, foreseeability is key. To induce compliance to the law, the regulator may decide to reward the best regulatory entities, which have fully complied with regulations. Another option might be to establish insurance schemes that would cover AI systems. The insurance companies will establish conditions for covering certain AI apps and services.

48 Lessig, *supra* note 33; *see also,* in this book, a chapter by Peter Wahlgren (2021).

49 *See* AIHLEG, *supra* note 2 at 21–22.

50 *See* Koops, *supra* note 44.

The proactive approach seems to have been endorsed by AIHLEG because it supports trustworthy AI. The AIHLEG has thus emphasized on education and awareness (information supply) and multi-stakeholder engagement, including design teams and consumer representatives.[51]

Last but not least is the *evolutionary approach*. This presents a shift from behavior norms to duty of care norms, liability rules,[52] regulatory sandboxes, and self-replicating rules,[53] i.e., intelligent laws for intelligent machines. In the future, it might be possible to think of self-replicating laws. These should be designed to regulate AI systems. As AI is evolving, there should also be research into intelligent (autonomous) laws. If this comes to fruition, it will be a fusion of code is law and the evolutionary approach. It can be foreseen that there will be questions as to the role of the legislature and existing law enforcement agencies.

The evolutionary approach also includes adoption of cost-benefit analyses of laws and regulations. This has developed into what is famously referred to as Better Regulation, which entails legislative and regulatory impact assessments. The approach is relevant in the regulation of AI, and there is evidence already of its application, e.g., GDPR requires privacy impact assessments (Article 35 of GDPR).[54] Somewhat related to the technology impact assessment is the conformity assessment introduced by the proposed EU AI Regulation.[55]

51 AIHLEG, *supra* note 2 at 23.

52 *See*, for example, the virtual legal duty of care as analyzed by Conrad Nyamutata, *Childhood in the Digital Age: A Social, Cultural and Legal Analysis of the UK's Proposed Virtual Legal Duty of Care*, 27 Int'l J.L. & Info. Tech. 311–338 (2019); another example of the duty of care norm is Volvo's guarantee to cover any AI system errors in their cars. For details *see* William D. Eggers, Mike Turley & Pankaj Kishnani, *The Future of Regulation: Principles for Regulating Emerging Technologies*, Deloitte Insights (June 19, 2018), https://www2.deloitte.com/us/en/insights/industry/public-sector/future-of-regulation/regulating-emerging-technology.html (last accessed May 30, 2021).

53 *See, e.g.*, Gunther Teubner's law as autopoietic system: Gunther Teubner, Law as an Autopoietic System (1993); Gunther Teubner, *Substantive and Reflexive Elements in Modern Law*, 17 L. & Society Rev. 239–286 (1983); Article 53 of the Proposed EU AI Regulation COM (2021) 206 (provides for AI regulatory sandboxes).

54 *See also* Hoffmann-Riem, *supra* note 8 at 33. *See* the Proposed EU AI Regulation, COM (2021) 206 final.

55 Article 19 of the proposed EU AI Regulation deals with conformity assessment. *See also* Article 43 of the same regulation, providing for a conformity assessment procedure;

Under the proposed law, AI providers will be required to apply to national authorities for a conformity assessment. In that process, a particular AI will be assessed as regards its conformity with the law. Technology impact assessments are not a new regulatory strategy. In the USA, they have been used since 1960s.[56] A problem with impact assessments is that they are costly. Moreover, technology development may be spontaneous and disruptive.[57]

If the legislature and regulators embrace the LTs as briefly discussed above, so-called trustworthy AI may be developed. The AIHLEG Guidelines for Trustworthy AI state that there are three features of trustworthy AI: it must be lawful, ethical, and robust. However, the AIHLEG Guidelines focus on ethics (human dignity – beyond rights and law) and robustness (safety and security). These have not been specifically discussed in this chapter.

3 What are the applications?

To discuss how LTs may be applied to the regulation of AI, this is exemplified using unmanned aircrafts (drones), eHealth apps and services, LAWS, and autonomous vehicles. The regulation of AI may focus on design, development, and application of AI systems. Regulatory challenges may arise either in the development or use or the results of use or operation of a particular AI system. In a work of limited space, it is difficult to make a detailed treatment of the examples selected above, therefore the below is a mere sketch.

(1) Unmanned aircrafts

While traditional civil aviation laws may be unfit to regulate unmanned aircrafts, other techniques such as contract-based liability rules and duty of care norms may be operative in the regulation of drones. Nevertheless, certain regulators have recently

---

Article 16 imposes obligation on providers of high-risk AI including to do conformity assessments and to have a quality management system in place. The data specified in Article 13 of the proposed EU AI Regulation COM (2021) 206 will be used for privacy impact assessments.

56 *See* Emilio Q. Daddario, *Technology Assessment – A Legislative View*, 36 Geo. Wash. L. Rev. 1044–1059 (1968).

57 Ubena, *supra* note 33.

resorted to TLTs for authorization, licensing, and registration to regulate drones.[58] Consequently, operating a drone (regardless of size/weight) requires a license in some countries, such as Tanzania.[59] These requirements were made by the regulator after considering how drones operate and the emerging risks in terms of security and safety. Nonetheless, the license and registration regime might have not taken into consideration issues of liability and the relationship between the drone manufacturer and the operator on the one hand, and the relation of the operator, the software service provider and the internet service provider on the other. If a drone has a malfunction and an accident occurs, who should be responsible? Perhaps in such situation, contracts, code is law, or proactive and evolutionary approaches may come into play.

(2) eHealth apps and services

The command and control approach may apply in defining the rights and obligations of eHealth actors, including app developers, service providers and patients. However, it is possible to use complementary approaches, such as codes of conduct developed for AI-based eHealth systems. An analogy may be drawn from GDPR (Articles 40–41) on codes of conduct for data processing companies. The certification of AI-dependent eHealth services may also be a viable strategy. It can be foreseen that there will be authorities for certifying AI systems for eHealth in the future. From a general perspective, similar regimes have already been embraced in the proposed EU AI Regulation.[60] This is likened to the establishment of regimes under GDPR (Articles 42–43), such as data protection certifications, seals, and marks to enhance trust and indicating quality

58 The Canadian Transport Authority has published requirements for flying drones. They are part of the old laws, i.e., Civil Aviation Regulations for drones. *See* Transport Canada, *Flying your drone safely and legally*, Government of Canada, https://www.tc.gc.ca/en/services/aviation/drone-safety/flying-drone-safely-legally.html (last accessed May 30, 2021).

59 *See* Tanzania Civil Aviation (Remotely Piloted Aircraft Systems) Regulations, Government Notice No. 758 (2018).

60 The certification is also covered in Article 16 of the proposed EU AI Regulation COM (2021) 206 on ethical principles for the development, deployment and use of artificial intelligence, robotics, and related technologies.

assurance.[61] It would also be possible to put into place liability rules for eHealth services. However, these require understanding of the actors involved in the AI-based eHealth service value chain. Without such understanding, blame may be placed on the wrong party.

(3) Lethal autonomous weapons systems

The emergence of lethal autonomous weapons systems (LAWS) that take lives without human intervention is not fiction anymore. They are reality. From the traditional law perspective, LAWS could be banned,[62] but that could stifle innovation. Moreover, LAWS may be regulated through the adoption of an international convention on disarmament.[63] However, if the convention does not specifically mention LAWS, this may be difficult. It may also be a challenge to attribute a certain attack to a particular state, especially if the attack was due to dysfunctional LAWS. That may be complicated if LAWS are outsourced from, say, country B, the country using the system is country A, and the attack is on country C. The situation may further be complicated as certain LAWS are self-executing programs. Therefore, the traditional law ought to be complemented by embedded technical solutions, and evolutionary and proactive approaches. This might make it possible to have duty of care norms, liability rules, impact assessments and information supply, especially regarding the algorithm, to enhance the transparency of the system.

61 Nikolaus Marsch, *Artificial Intelligence and the Fundamental Right to Data Protection: Opening the Door for Technological Innovation and Innovative Protection*, in Regulating Artificial Intelligence at 48–49 (Thomas Wischmeyer & Tim Rademacher (eds.), 2020).

62 Kenneth Anderson and Matthew C. Waxman, *Law and Ethics for Autonomous Weapon Systems: Why a Ban Won't Work and How the Laws of War Can*. American University, WCL Research Paper 2013-11, Columbia Public Law Research Paper 13-351 (2013), available at SSRN: https://ssrn.com/abstract=2250126 or http://dx.doi.org/10.2139/ssrn.2250126. More sources on LAWS are available at https://www.un.org/disarmament/the-convention-on-certain-conventional-weapons/background-on-laws-in-the-ccw/. *See also* UNODA, note 63.

63 *See* for example UN Convention on Certain Conventional Weapons of 1980 (United Nations, *Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons Which May be Deemed to be Excessively Injurious or to Have Indiscriminate Effects (and Protocols) (As Amended on 21 December 2001)*, 10 October 1980, 1342 UNTS 137.) Its application to AI may be limited because AI weapons may be regarded as non-conventional weapons. However, that may be a matter for interpretation.

(4) Autonomous vehicles

Already, car manufacturers like Volvo have offered assurances that they will be responsible in case of errors in their driverless cars. Under such circumstances, one might consider adopting insurance for autonomous vehicles. The insurance regime uses the proactive approach, where the purpose is to foresee dangers and attempt to address them before they actually occur. The code is law approach could also be adopted, as embedded solutions may be included in the design and development of autonomous cars. A car may be designed in such a way that it can, through algorithms or sensors, calculate the likelihood of the occurrence of an accident such as a collision, so the engine can stop instantly. This would be possible if the roads were AI-friendly and all the vehicles on the roads had sensors that enabled communication with others using the same roads. An old, but good, example of how this can be developed is seen in Breathalyzers, the Saab AlcoKey, and similar devices and apps used for testing a driver's blood alcohol content levels before a car can be started. If the alcohol content is above the prescribed limit, the car engine will not start.

4 A human-centric AI?

LTs can support realization of a human-centric AI because the techniques are lawful, legitimate, necessary in democratic states and proportionate.[64] Furthermore, they are used in conformity with fundamental rights. For that reason, the techniques can help regulators achieve what other experts, such as AIHLEG, have suggested and advocated for, i.e., human-centric AI (ideally, AI regulation should be human-centric).

According to the EU AIHLEG, AI ought to be ethical and embrace societal values. The LTs provide regulators with a regulatory toolbox that blends TLTs and NRC, creating a multimodal approach that may support human-centric AI. However, human-centric AI cannot be achieved without understanding the regulatory environment. It requires a constant monitoring and evaluation process. The involvement of expert groups to monitor how AI evolves is paramount.

64 For details on the pre-conditions of legislative techniques, *see* Ubena, *supra* note 4 at 89–97.

Furthermore, it should be remembered that regulation is a living process and legislation is a scarce resource. AI regulation should entail multi-stakeholder engagement, involving both state and non-state actors. Above all, it should be multi-layered – nationally and globally.[65]

5 Conclusion

The chapter has shown that it is possible to regulate AI only if the right regulatory tools are in the hands of the legislature and regulators. LTs have been depicted as a possible starting point for regulation of AI, as they have the potential to match the development of AI. Since the techniques have some shortcomings, the legislature ought to use various complementary techniques that might be useful and suitable in the regulation of AI. This chapter has not covered the link between AI and the rule of law or fundamental rights. Nor has the transnational nature of AI regulation been examined. Additionally, the emergence of intelligent laws to match AI development was mentioned only in passing. LAWS were treated superficially. Furthermore, the role of standard-setting organizations in the regulation of AI was not covered. These are interesting areas for future studies. Examination of these areas will promote the understanding of the multifaceted relationship between law and AI.

65 *See* Urs Gasser, *The Impact of Law and Regulation on Digital Technologies in Thailand*, August 23, 2017, presentation in Bangkok; Ubena, *supra* note 33.

Regulation of AI: Problems and Options

HÅKAN HYDÉN

Lack of experience and knowledge creates regulatory problems

In the paradigmatic shift from analogue to digital technologies (data communication, storage and AI) and from an industrial to a digital society, normativity changes from being connected to the use of technology to becoming an integral part of the technology as such[1]. This has both positive and negative effects. AI is used for decision-making, learning and performing tasks based on data, where the data are often complex, ambivalent and difficult to interpret. Areas that require some form of stability, clear goals, measurability and long-term vision can be expected to use AI in the future. Examples include fully or partially automated banks (both private and national banks) and automated systems for diagnostics and treatment (for example for diabetics). Some state bodies, such as customs, the police, fire brigades, roads and transport authorities, could be more data-driven and use AI to improve their ability to make the right decisions, for instance when optimizing budgets, maintenance and expansion and to develop the kind of skills and measures needed to achieve political goals. This expected development would not affect regulation of AI as such.

In the industrial era, development was a matter of prolongation within the same technological area – mechanics – through many

1 Cf. Langdon Winner, *Do Artifacts Have Politics?* 109 Daedalus 121–136 (1980).

small, incremental changes, rather than a qualitative shift. This goes for industrial production of cars, infrastructure, etc. The development and refinement of technology has, however, had external effects, which were all addressed using one and the same governance strategy, namely intervening law and controlling public authorities. As examples, legislation on consumer and environmental protection can be mentioned. However, when mechanical technology created through physical production turns into digital technology working in tandem with virtual reality, there are no reference points. This becomes a problem for regulators, as adequate knowledge of what is to be regulated is a prerequisite for regulation.

In the case of AI, this makes the normative problems seem philosophical, lacking any solid answers. We can only wait to gain more experience[2]. This turns the problems into empirical questions and the answers become socio-legal, i.e., advanced practice will provide us with various, tentative, practical solutions, which can lay the foundation for normative assessments. A common denominator is uncertainty regarding consequences, combined with the ambivalence that characterizes policies. Therefore, regulatory problems are, in these cases, mostly referred to and discussed in terms of ethics[3]. AI represents a new regulatory phenomenon. The legal principles will not be different, but the substratum (i.e., reality) changes, which makes it necessary to adjust and reformulate legal regulations.

As an example, we can refer to self-driving cars and a comparison between the fatal accident with the first automobile in the late 19th century and the first fatal accidents with autonomous cars 120 years later. The difference is the driver – or the lack of driver, in the latter case – but the legal problems are otherwise the same[4]. Regardless,

2 Fumio Shimpo, *The Principal Japanese AI and Robot Strategy Towards Establishing Basic Principles,* in Research Handbook on the Law of Artificial Intelligence (Woodrow Barfield & Ugo Pagallo (eds.), 2018).

3 *Ethics Guidelines for Trustworthy AI*, High-Level Expert Group on Artificial Intelligence (HLEG) set up by the European Commission 2018. The Trustworthy AI assessment list presented in this document from 2019 has undergone a piloting phase among stakeholders to gather practical feedback. Based on the feedback received, the AIHLEG presented the final Assessment List for Trustworthy AI (ALTAI) in July 2020. See also Larsson, S. (2020). *On the Governance of Artificial Intelligence through Ethics Guidelines*. Asian Journal of Law and Society, 7(3), 437-451. https://doi.org/10.1017/als.2020.19.

4 Jannice Käll, *Governing Smart Spaces Through Autonomous Vehicles*, *in* Smart Urban Mobility. Law, Regulation, and Policy, Michèle Finck et al. (eds.), 2020).

liability is the main problem related to negligence, something elaborated on below. The race by automakers and technology firms to develop self-driving cars has been fueled by the belief that computers can operate a vehicle more safely than human drivers. However, that view has partly been called in question after two fatal accidents involving self-driving cars. The first case was that of a driver of a Tesla Model S electric sedan killed in an accident when the car was in self-driving mode. Federal US regulators are in the process of setting guidelines for autonomous vehicles. In a statement, the National Highway Traffic Safety Administration said that reports indicated that the crash occurred when a tractor-trailer made a left turn in front of the Tesla and the car failed to apply the brakes. The second recorded case is a pedestrian fatality involving a self-driving car, following a collision that occurred late in the evening of March 18, 2018. A woman was pushing a bicycle across a four-lane road in Tempe, Arizona, United States, when she was struck by an Uber test vehicle, which was operating in self-driving mode with a human backup driver sitting in the driver's seat. The problem was related to how the car detected objects in the road.

These cases brings to the fore safety regulation in connection with these specific AI incidents. The second case also highlights the question of liability. Who is responsible if you remove the driver? Is it the producer of the vehicle, the entity providing training data used for learning in the vehicle self-driving system, the person who uses the vehicle, the driver, or the owner of the vehicle? As of April 2019, 33 states in the US had enacted legislation pertaining to autonomous vehicles[5]. In Sweden, a law has been proposed (SOU 2018:16) which introduces new rules for drivers, owners and legal persons as regards the conditions for criminal liability. Among other things, there is a discussion of introducing vicarious liability and strict liability in relation to AI. Since neither statutory nor common-law jurisdictions accept AI's status as a legal person, AI cannot be a principal or agent[6].

5 National Conference of State Legislatures, *Autonomous Vehicles – Self-Driving Vehicles Enacted Legislation*, http://www.ncsl.org/research/transportation/autonomous-vehicles-legislation.aspx (last visited May 1, 2021).

6 Woodrow Barfield, *Towards a Law of Artificial Intelligence*, in Research Handbook on the Law of Artificial Intelligence 37 (Woodrow Barfield & Ugo Pagallo (eds.), 2018); David C. Vladeck, *Machines Without Principals: Liability Rules and Artificial Intelligence*, 89 Wash. L. Rev. 117–150, 127 (2014).

One can compare these fatal accidents with that of Bridget Driscoll, the first pedestrian to be killed by an automobile – in the United Kingdom in 1896[7]. The situations are similar. In 1896, it was a question of introducing a new means of transportation based on inventions in mechanics. This can be compared to the Uber accident, which was related to the introduction of a digitally innovative means of transportation. The car crash in Arizona seemed as startling — and perhaps as likely, given the pioneering technology of autonomous driving — as the car crash that killed Driscoll in southeast London on a summer's day more than a century ago. This case can serve to illustrate how technology affects society. As in many motor vehicle accidents, there were conflicting accounts of what happened on Aug. 17, 1896[8]. Testimonies focused on the vehicle's speed, the driver's abilities, and whether the public had been given enough warning about the demonstration vehicles[9]. There are lessons to be learned from similar situations and corresponding phases in past societal developments.

From both/and to either/or

Digital technologies give rise to a paradigmatic shift in regulation strategies. In the industrial society, regulation was demanded mainly due to the external effects of production, distribution and consumption. Production leads, among other things, to overconsumption and energy waste – matters where the political system has intervened by way of law. While production is desirable, the negative external effects it has must be minimized, which calls for compromises. Thus, regulation must be based on compromises between contradictory goals. You can eat the cake and have it too. Such regulation requires public authorities that implement and perform supervision.

7 Fredrick Kunkle, *Fatal Crash With Self-Driving Car Was A First — Like Bridget Driscoll's Was 121 Years Ago With One of the First Cars*, Washington Post (March 22, 2018), https://www.washingtonpost.com/news/tripping/wp/2018/03/22/fatal-crash-with-self-driving-car-was-a-first-like-bridget-driscolls-was-121-years-ago-with-one-of-the-first-cars/ (last visited May 1, 2021).

8 At the time, the Anglo-French Motor Carriage Co. was demonstrating the performance of three imported cars in the Dolphin Terrace, an area behind the Crystal Palace.

9 Florence Ashmore, a domestic servant who had witnessed the crash, said the vehicle had come on "at a tremendous pace, in fact, like a fire engine."

This can be seen to characterize the industrial world of today as regards consumer protection, environmental protection, etc.

In the digital era, the nature of the problems related to societal development has changed. New technologies create hitherto unknown possibilities. However, we can never predict the future, only anticipate it. The discipline of anticipation[10] is based on two prerequisites: prolongation of trends in society and pattern recognition. How fruitful the method is will depend on the relevance of the identified trends in society[11]. The problem is that we tend to think in straight lines. When we imagine how the world will change in the 21 st century, we just take the progress made in the 20th century and add it to the year 2000. Linear thinking is the most intuitive way of imagining the future, but we should be thinking exponentially[12]. According to the McKinsey Global Institute, the AI Revolution is "happening ten times faster and at 300 times the scale, or [with] roughly 3,000 times the impact" compared with the Industrial Revolution[13]. However, our experiences blind us to the future[14]. They prevent us from seeing that we are heading toward a completely new horizon. The regulation problem becomes a question of *either/or* instead of *both/and*. Regulation no longer aims at controlling external effects, as in the industrial mode of production; rather, a choice has to be made between alternative areas of application. The question to be asked in relation to the digital production is for what purposes we should accept the use of the new technology. For instance, is genetic modification desirable or should it be avoided?

10 Riel Miller, Roberto Poli & Pierre Rossel, *The Discipline of Anticipation: Exploring Key Issues*, in Transforming the Future (Riel Miller (ed.), 2018).

11 For an application of the theory, *see* Håkan Hydén, *Social Cohesion and the Anticipated Fall of the Welfare State*, 5 Ann. Soc. Sci. Manage. Stud. (2020).

12 Tim Urban, *The AI Revolution: The Road to Superintelligence*, Wait But Why (January 22, 2015), https://waitbutwhy.com/2015/01/artificial-intelligence-revolution-1.html (last visited May 1, 2021).

13 Richard Dobbs, James Manyika & Jonathan Woetzel, *The Four Global Forces Breaking All the Trends*, McKinsey Global Institute (April 2015), https://www.mckinsey.com/business-functions/strategy-and-corporate-finance/our-insights/the-four-global-forces-breaking-all-the-trends# (last visited May 1, 2021).

14 According to Urban, we base our ideas about the world on our personal experience, and that experience has ingrained the rate of growth of the recent past in our heads as "the way things happen."

New technologies that emerge are valued in the terms of the old. In our times, as we face new technological developments based on digitization, we have no previous experience to fall back on. We are, quite simply, facing hitherto unknown problems. This has previously been seen in the biotechnology field in connection with, among other things, genetic engineering. There is a knee-jerk reflex to expect new legislation to control such new phenomena. To the extent that new technology is abused, such abuse is already covered under existing criminal and civil law. Still, the efforts of large-scale society to control and subordinate everything lead to demands that new activities be controlled and subordinated. This has led to the introduction of regulations on genetic engineering in Chapter 13 of the Swedish Environmental Code. The regulations apply to genetically modified organisms (GMOs) being used in a laboratory environment, being released in a natural environment and to products containing GMOs being introduced on the market. The regulations have been introduced in environmental legislation, which has previously mainly addressed environmental protection, with the aim of sustainable development. However, with regard to GMOs, the purpose of the regulations is primarily to ensure that specific ethical considerations are taken into account. This is tested by the competent authority – the Swedish Board of Agriculture in the case of agricultural products. According to Chapter 13, Section 12 of the Environmental Code, a license is required to carry out a deliberate release of GMOs into a natural environment or when introducing GMO products on the market. Such license may only be granted if the activity is ethically justifiable. Additionally, a special committee, the Swedish Gene Technology Advisory Board, must submit an opinion. So far, these opinions and decisions have led to activities routinely being considered ethically justifiable. The law and legislative history are somewhat vague with regard to what would make a release or product introduction not ethically justifiable.

Surveillance may be another illustrative case in point. Algorithm-based facial recognition and monitoring of deviating behaviors are common today in both open and closed ecosystems. Such technology creates a number of new opportunities and can be used

for multiple purposes[15]. Examples include privately owned cameras that recognize family and friends but alert the owner to unknown visitors, and cameras used by customs officers and the police in public places to monitor and, in some cases, track people.

In the digital era, societal problems are more a matter of choosing between future options than compromises within the framework of an alternative. We are faced with a qualitatively new regulatory phenomenon as a result of AI, with surveillance serving as a prime example. It is much like the internet itself. The basic principle of the internet is that it should be open and free for everyone. Lawrence Lessig discusses "the norm of open code"[16]. This view of the use of the internet stands in contrast to another, which deals with surveillance and limitations of the internet for various purposes. The technological development behind the emergence of the digital society tends to be adopted and used within the framework of the existing, pre-digital society's logic and power centers, at least in a first stage. Legal regulations, at both a national and an international level, such as the European Convention on Human Rights, protect us from privacy violations on the part of the state and public authorities. These rules are often not applicable in cases of violations by large private corporations, which are the main sources of threats in an AI context.

The main argument for introducing surveillance is to provide security for ordinary people. There are many tools in our daily lives which help individuals protect their property and privacy. This is regarded as a positive effect of technological, digital developments. However, when such tools end up in the wrong hands, they can be turned into something evil. This is already taking place, to a large extent.

From ex post to ex ante regulation

In this situation, a process of trial and error arises. Since we are not certain of the potentials and consequences of new technology, it has to be developed with ethics kept in mind. The use of new technology appears to cause a problem of values, rather than practical

15 In China, this kind of surveillance was used during the COVID-19 pandemic, to ensure that people obeyed outdoor restrictions.

16 Lawrence Lessig, Code: Version 2.0 (2006).

issues. Discourses based on values are the forerunners of legal regulation. Ethical and political problems lead to the question of deciding where to draw the line in various activities. What kind of outcome will we accept? This creates a dilemma when we lack experience of the activities in question. During the industrial era, external effects functioned as triggering factors for interventions, a strategy based on evaluations *ex post*, which is problematic in the context of phenomena like AI and algorithms. We do not yet know which leg to stand on. This was reflected in a survey conducted in 2019 called "AI through the eyes of the consumers in the Nordic countries"[17]. In response to the question "To what extent do you think AI would make better/worse and more/less unbiased decisions than humans?", around half of the respondents thought that AI would make just as good/bad decisions, in some cases even much better decisions. The strongest support for AI is found in areas like the industrial sector, banking, accountancy, and public government, while in areas based on human-to-human interaction, such as healthcare and legal consultancy, people had greater trust in other humans.

As regards AI, regulations have to be based on ex ante considerations, at least as long as we lack knowledge about the normative consequences. In general, new technology emerges without political decisions and needs no support from the legal system[18]. Quite the opposite, it often requires de-regulation. The growing digital technology is a case in point[19]. It is self-promoting in a way that might collide with laws in the affected legal fields. The state can stimulate and promote certain solutions by setting up special zones for empirical testing and development. For example, in the field of AI, the Japanese government has initiated a kind of living lab,

17 The survey was conducted by YouGov on behalf of Tieto Sweden Ltd and was based on an analysis of 3,659 computer-assisted web interviews with Swedes, Norwegians and Finns aged 18+ years on 10–12 April 2019. Data were weighted based on respondents' gender, age, and geography in order to be representative for the population.

18 Mireille Hildebrandt, Smart Technologies and the End(s) of Law: Novel Entanglements of Law and Technology (2015).

19 Research Handbook on the Law of Artificial Intelligence xxv (Woodrow Barfield & Ugo Pagallo (eds.), 2018).

called Tokku[20]. In the field of autonomous vehicles, several EU countries have endorsed similar experiments. Sweden has sponsored the world's largest-scale autonomous driving pilot project, in which self-driving cars use public roads under everyday driving conditions.

Law is actualized primarily for preventive reasons, as a reaction to the negative aspects of new technology. These negative aspects are not new and can therefore be combatted with existing legal means. The difference is rather what causes the problems. Whenever a need for legislative action or laws related to AI is suggested, this is primarily in relation to damages caused by AI[21]. The new technology, with its changes to the substratum, will be subsumed into existing legal paradigms and adopted into existing legal principles and rules[22]. Karl Renner, in his study of the institutions of private property and its societal functions, pointed out that the legal institutes might be the same, but they had been combined differently during specific phases of legal development from Roman law and through to the modern law era. The first step in a transitional process is to make analogies[23]. The question that should be asked is "what is similar" to the issue at stake for the legal decision-making process.

Common law (also referred to as case law) has an advantage compared with statutory law since it is based on decision-making on the part of judges, and the legal doctrines are established by judicial precedent rather than by statute. This forces common law systems to confront new societal phenomena at a much earlier stage than statutory-based legal systems, like the continental European system.[24] In the statutory legal systems, a legal matter has to be approved by the

20 *Id.,* p. xxvi. The unique *"Tokku" Special Zone for Robotics Empirical Testing and Development* (RT special zone) originated in Japan. Since 2003, the world's first RT special zone had been established in Fukuoka Prefecture, Fukuoka City and Kitakyushu City. The feasibility of bipedal humanoid robots on public roads was studied there from 2004 to 2007. These were the world's first public road tests for bipedal robots.

21 *Id.* at xxiii.

22 Karl Renner, The Institutions of Private Law and Their Social Functions (1949).

23 Curtis E.A. Karnow, *Foreword*, in Research Handbook on the Law of Artificial Intelligence xxi (Woodrow Barfield & Ugo Pagallo (eds.), 2018).

24 The legal material regarding AI therefore to a large extent consists of cases brought before US courts.

legislative body, i.e., the political system. This means that the statutory legal system is characterized by greater inertia than the common law system[25]. Judges have to take a stand in a legal matter, even if the problem is unknown and without precedent, while decision-making, especially in democratic processes, takes a long time for politicians, who act based on views and opinions. Formulating a political will, which requires experience of a new phenomenon, takes time.

AI and algorithms lead to interpretation problems and legal policy considerations when the substratum of law undergoes changes as a consequence of the new conditions for regulation[26]. The new conditions which judges and/or legislators face are primarily an effect of three factors[27]: AI's autonomic functioning, the complexity and transparency problem and, lastly, the need for big data. Some areas have already been subject to legal regulation due to AI, such as data protection, security and liability rules, robots, antitrust law, and consumer protection[28]. These are areas where problems already existed, but were scaled up under the influence of digitization and therefore required certain precautionary measures. Another matter is that there are new causes of discrimination as an indirect effect of machine learning and big data.

25 In my view, this is a question of scaling. The digital technology faces problems which have been there before, even if the digital technology accentuates the problem. Horwitz, Morton J. (1994) (The transformation of American law, 1870-1960: the crisis of legal orthodoxy. New ed. New York: Oxford University Press) traces the development of common law followed by statutory law, as a function of economic forces. See also Karnow, supra note 23 at xix.

26 Barfield & Pagallo, *supra* note 19.

27 Daniel Westman, *Den fjärde industriella revolutionen – en immaterialrättslig introduktion*, 1 Nordiskt Immateriellt Rättsskydd 131 (2019).

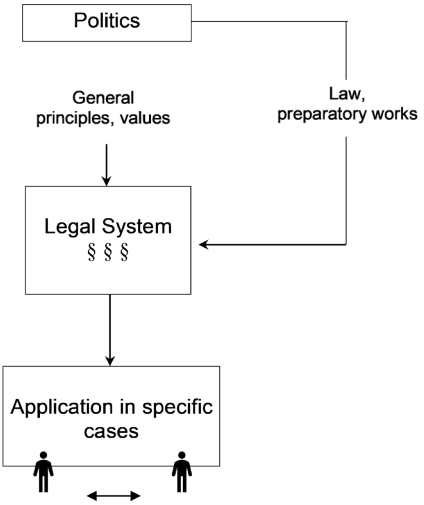
28 I will not delve into how law in itself can use AI for different purposes, such as Correctional Offender Management Profiling for Alternative Sanctions or COMPAS. COMPAS is a case management and decision support tool developed and owned by a private company (now Equivant) used by U.S. courts to assess the likelihood of a defendant becoming a recidivist based on scales using behavioral and psychological big data (https://en.wikipedia.org/wiki/COMPAS).

Different orders of normativity

The main regulatory problem relates to identification of the normative consequences of AI. There is, however, a crucial difference between algorithms in a technical sense and algorithms in a social science perspective. Both are normative, but they cover different fields of knowledge. One parallel is legal norms. They can be understood from a strictly legal point of view, telling us about the correct interpretation and application of a legal rule, as an instruction on how to act or how to judge in a certain situation. However, legal norms also have a broader scope in a social science perspective. Legal norms are not neutral, as they affect societal functions and have consequences for society. I claim that the point is that there are different orders of normativity: the first is related to the algorithm as a technical instruction and the second to the consequences springing from the first order. To illustrate with an example from the legal field: it is one thing to know when a person should be sentenced to imprisonment and another to understand what this means for society, the perpetrator and the victims of the crime. These are distinct spheres of knowledge, which require different methodological approaches: the legal dogmatic approach, on the one hand, and the social science perspective within sociology of law and criminology, on the other. The normativity layers associated with algorithms are special and understanding the second order calls for a separate concept, what I call *algo norms.* They are an indirect effect of the algorithms and it is this indirect effect which is of interest from a sociology of law perspective. Thus, *algo norms* are related to the societal consequences, which follow from the use of algorithms in different respects.

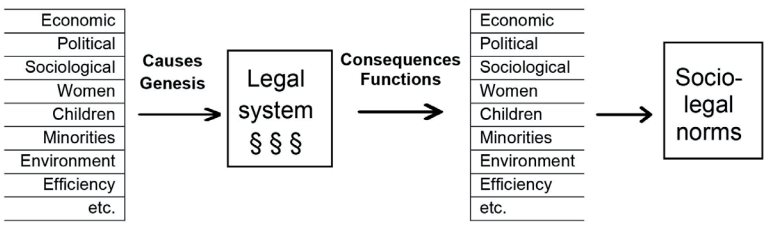
Let me continue with the parallel between law and AI. The similarities and differences between legal norms and *algo norms* can be illustrated graphically. If we start with the legal knowledge field, the following figure might give an understanding:

Figure 1. Legal decision-making from a legal dogmatic perspective.



The legal dogmatic perspective can be illustrated vertically, since it – as an ideal type (in a Weberian sense) – is based on the logics of subsumption and deduction[29]. This involves technical application of normative standpoints in law to factual situations, which may require more or less sophisticated reasoning. We can also extend the legal knowledge field to include socio-legal aspects covering the causes and consequences of law, i.e., looking at the genesis and functions of law, which is the focus for sociology of law[30]. See below:

Figure 2. The legal system from a socio-legal perspective.

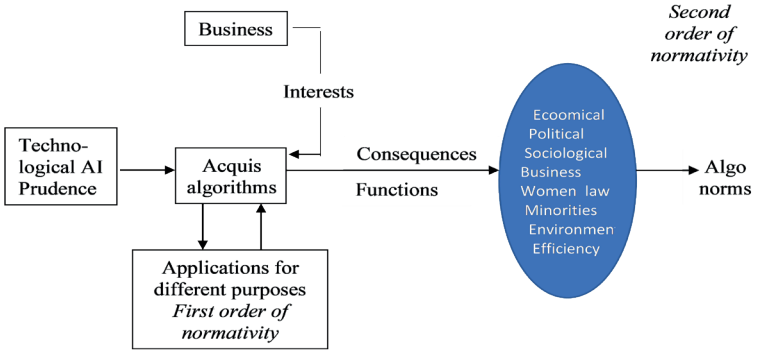


29 Håkan Hydén & Therese Hydén, Rättsregler: En Introduktion till Juridiken (2019).

30 *Contributions in Sociology of Law: Remarks From a Swedish Horizon*, 29 Lund Studies in Sociology of Law (Håkan Hydén & Per Wickenberg (eds.), 2008).

This horizontal problem area represents something other than the legal dogmatic knowledge field, but is of great relevance for understanding law. It is another view of law, which is not regarded as relevant in the legal dogmatic perspective. Sociology of law has not, thus far, invented a concept which covers the normativities related to the genesis and consequences of law when applied to and confronted with societal realities. The concept of law in action[31] is not adequate, nor is the concept of living law[32] or the distinction between the manifest and latent functions of law[33]. We are looking for something else, which has to do with the indirect social consequences of using AI. If we look for a parallel to the digital world, we can translate the two illustrations into the *algo norm* context. We then get the following figure:

Figure 3. AI's normative context.



Algo norms, the indirect effect of AI

Algo norms can be regarded as a subcategory of technical norms. The theoretical perspective underlying the concept of *algo norms* is based on norm science theory and method[34]. Norm science is about iden-

31 Roscoe Pound, *Law in Books and Law in Action*, 44 Am. L. Rev. 12 (1910).

32 Eugen Ehrlich, Fundamental principles of the sociology of law (2002).

33 Robert K. Merton, Social Theory and Social Structure: Toward the Codification of Theory and Research (1949).

34 Håkan Hydén, *Looking at the World Through the Lenses of Norms. Nine Reasons for Norms: A Plea for Norm Science*, in Understanding Law in Society. Developments

tifying and understanding the driving forces behind human action at a societal level[35]. The study of norms tends to be divided into two perspectives, one descriptive and one injunctive[36]. Banakar also uses the parallel terminology of "external" and "internal" perspectives on norms. However, there is a third possible understanding of the norm concept, often ignored in the social sciences: the analytical perspective[37]. Here, the norm concept and the empirical study of norms help us understand causalities underlying human behavior at a collective level. Through the study of norms, human motives for collective action can be captured. This approach goes beyond Max Weber's *Verstehen* method[38]. Weber was a methodological individualist and suggested that we can only understand social phenomena and historical processes by studying how individuals experience the world and what individuals find to be meaningful. By dissecting existing norms in a descriptive way, it is possible to get hold of the preferences and motives that underlie human behavior at a collective level.

The identification of *algo norms* as the normative outcome of AI has its own challenges, since the normativity is an implicit external effect, not an explicit one. The primary objective of the algorithms is not to produce a set of norms. Rather, they are hidden effects that need to be made visible through studying their societal effects. The normative dimension is hidden behind cognitively based instructions on how to act. It should be noted that this is not the same as the problem with the so-called black box. Locating the regulatory problem within the black box makes us focus on how AI is constructed and what the technicians had in mind. Thus, from a social science perspective, transparency becomes the main obstacle to knowledge. The problem from a regulatory point of view, which

---

in Socio-legal Studies (Knut Papendorf, Stefan Machura & Kristian Andenaes (eds.), 2011).

35 *Id.*

36 Banakar, Rez, Normativity in Legal Sociology: Methodological Reflections on Law and Regulation In Late Modernity 216 (2014).

37 Håkan Hydén, Sociology of Law as the Science of Norms, Routledge Publ., 2022.

38 Max Weber, The Methodology of the Social Sciences (1949).

I want to highlight, lies outside "the black box" and is a question of what happens when AI is applied in society.

In order to dissect norms in terms of motives, we must distinguish between three dimensions of the norm: (1) Will and values, (2) Knowledge and cognition, and (3) Systems and possibilities[39]. Typically, social norms are created based on human desires and values, which require knowledge for implementation, including cognitive references to the norm's addressees. The outcome of a norm application is ultimately dependent on the possibilities of doing what the norm prescribes. Systems that humans have created for various purposes set the limits of these possibilities. *Algo norms* are different from social norms[40] in that their genesis is related to new knowledge and digitization, thus generating their own systems with different purposes. These systems, in turn, influence desires and values. The desires and normative consequences are subordinate to the knowledge and the systemic effects generated.

Algo norms emerge when algorithms meet and collide with the surrounding society, i.e., the second order of normativity. Different consequences arise when technological solutions and design are applied in reality; some can be seen as intended, but many are unintentional. They are external effects of the algorithms. Thus, there are two kinds of causality. First, the algorithm, as a technical instruction, performs a certain service. This service is meant to fulfil a specific purpose, which goes beyond the mere technical aspects of algorithms. The relation between the two steps is often invisible.

Algorithms as norms are unique. The normative consequences are embedded in the technology and determined by the design of the AI. The outcome is an empirical matter[41]. They are, from the perspective of the addressee, structurally conditioned and cannot be avoided. As technology historian Melvin Kranzberg (1986) expresses it in his *first law of technology*, algorithms are neither good nor bad;

39 Hydén, *supra* note 30.

40 Cristina Bicchieri, The Grammar of Society, The Nature and Dynamics of Social Norms (2006); Robert C. Ellickson, *The Evolution of Social Norms: Perspectives From the Legal Academy*, in Social Norms (Michael Hechter & Karl-Dieter Opp (eds.), 2001).

41 Carlos Alvarez-Pereira, *Disruptive Technologies, A Critical Yet Hopeful View*, 3 Cadmus (2017).

nor are they neutral[42]. What he suggests is that "technology's interaction with the social ecology is such that technical developments frequently have environmental, social, and human consequences that go far beyond the immediate purposes of the technical devices and practices themselves, and the same technology can have quite different results when introduced into different contexts or under different circumstances."

In this way, Kranzberg confirms the idea of a first and second order of normativity. The first is precise and related to technology, while the second is diversified and multi-normative. As Adrian Mackenzie has further observed, "[a]n algorithm selects and reinforces one ordering at the expense of others"[43], often in ways that were not intended or possible to foresee. *Algo norms*, therefore, are norms to which people are subordinated – in ways that lie largely outside their control. *Algo norms* are neither a matter of free will, nor one of coercion. The design of the technology and its normative implications are determined by people with technical expertise. In this perspective, engineers become our new norm-setters, at least as long as AI is logical and in the hands of humans, as opposed to being determined by the technology itself[44].

From a social science point of view, *algo norms* are problematic in two inter-connected ways, which affect the regulatory options. One is the democratic deficit that arises when norms are introduced into society, having been decided upon by technicians or by the system of algorithms itself. They are then neither the result of political decision-making in a democratic order nor an outcome of social or public discourses. This has consequences regarding the options for regulation within legal science. Sociology of law's knowledge interest is related to how decisions are made and with what normative implications, in order to make it possible to control the outcomes of AI. Even with the best intentions to create algorithms that make life better for people, the values and prejudices of those who feed the

42 Melvin Kranzberg, *Technology and History: "Kranzberg's Laws"*, 27 Tech. & Culture 544 (1986).

43 Adrian Mackenzie, Cutting Code: Software and Sociality 44 (2006).

44 Niels ten Oever's research focuses on how norms, such as human rights, get inscribed, resisted, and subverted in the internet infrastructure through its transnational governance. *See* Niels ten Oever, Wired Norms. Inscription, Resistance, and Subversion in the Governance of the Internet Infrastructure (2020).

algorithm with data and design the code will affect how the algorithms are constructed[45]. Furthermore, algorithms are to an increasing extent reproducing themselves. The opportunities for public accountability shrink and citizens face the risk of becoming captives of technical fixes over which they have little, if any, control.

The second related problem concerns manipulation in different respects, one of which relates to the market. *Algo norms* challenge the ideal role of the market as a tool for consumers to find goods and services. They confront us with a paradox. Our choices are determined by the algorithms and those who have programmed them in order to figure out what we like best and thus seem likely to want more of. We face a situation where the seller determines the content. Whenever someone uses the internet to buy products, view the news, access social media or browse the web, algorithms decide what they will find. This is an built-in effect of the technology – actually, it is its *raison d'être*. The market is in itself an algorithm (supply and demand meet in a computer system), but the actors are usually human beings; they interact with the market via computer screen, keyboard and mouse[46]. However, the *algo norms* are so seductive that we do not notice that information filters affect us. Not even the programmers are really aware of what is going on[47]. Tracing results from personalized searches, a website algorithm selectively guesses what information a user would like to have and encapsulates the user in a filter bubble[48]. As a result, users become separate from information that does not match their preferences or viewpoints, effectively isolating them in cultural and ideological bubbles[49]. The choices made

45 Lessig, *supra* note 16.

46 Donald Mackenzie, *A Sociology of Algorithms: High-Frequency Trading and the Shaping of Markets* (unpublished paper), https://uberty.org/wp-content/uploads/2015/11/mackenzie-algorithms.pdf.

47 A Swedish journalist and writer, Per Grankvist, has argued that algorithms appear in accordance with the same unwritten rules that have always applied to upper-class service staff. They should never draw attention, never make noise or be visible. Algorithms have learned what their master wants and provide these services without the master having to tell them to do so, http://pergrankvist.se/perspektiv.

48 Engin Bozdag, *Bias in Algorithmic Filtering and Personalization*, 15 Ethics & Info. Tech. 209–227 (2013).

49 Eli Pariser, The Filter Bubble: What the Internet Is Hiding from You (2011). This phenomenon reinforces the confusion and polarization, which are consequences of the transition in society from an industrial mode of production to a digital one.

by these algorithms are not transparent and it is difficult to foresee how they affect our worldviews and/or preferences.

Reactive or proactive regulation

The governance problem in relation to AI is a question of a proactive or reactive regulation strategy. In order to approach this problem, a new research agenda is required, which diverges from social science based on the industrial model of society. In the future – whether we like it or not – it seems that the preferences of society will not be determined only by politicians through law or by ordinary people through social norms. Instead, engineers and market players will have the strongest control. There are many indications of the economic influence on the development of AI[50].

The market forces are the main allies of AI. If one wanted to trace the driving forces behind AI development, the economic system would, as in almost any other societal issues, provide the answer. Is it possible to proactively influence this development? This would require research into AI, to make visible and articulate the driving forces, the contents of these norms and their hidden preferences; still, it would be hard to have any proactive influence. Since we, at least for the time being, lack knowledge about the new digital technology's effects in different societal respects, the simplest option is to stick to the strategy of trial and error, i.e., waiting to see what the consequences are and then making decisions about preventive actions. The options for regulation depend on collecting and systematizing experiences of AI's various consequences. Since we are dealing with regulation in relation to damages and negative aspects, we have reason to expect that certain events – such as scandals and problems of various kinds – will trigger action.

At the same time, we should be aware that once AI is unleashed, it may be too late to intervene. This calls for a proactive strategy. A unique feature of AI as a regulatory problem is its capacity of self-reproducing, even without human involvement. AI creates systems that not only reproduce and maintain themselves: they go a step further in being able to develop themselves, in a kind of autonomous process, and turn into something other than what they were. As a

50 Karnow, *supra* note 23.

consequence, we are caught in a dilemma between a desirable proactive strategy and an existing reactive strategy. We have to wait and react in order to figure out a proactive strategy. It seems that nobody has ownership of the matter of where technological development should lead us and what technology should be allowed to do. This brings us back to where this article started, with the lack of accumulated knowledge. We have no idea of or vision for what society we expect in the future and what we want to defend with law or even what institutions will be sustainable in the future.

Non-Asimov Explanations: Regulating AI Through Transparency

CHRIS REED, KERI GRIEMAN AND JOSEPH EARLY

Abstract

An important part of law and regulation is demanding explanations for actual and potential failures. We ask questions like: What happened (or might happen) to cause this failure? And why did (or might) it happen? These are disguised normative questions – they really ask what *ought to* have happened, and how the humans involved *ought to* have behaved.

If we ask the same questions about AI systems we run into two difficulties. The first is what might be described as the 'black box' problem, which lawyers have begun to investigate. Some modern AI systems are highly complex, so that even their makers might be unable to understand their workings fully, and thus answer the *what* and *why* questions. Technologists are beginning to work on this problem, aiming to use technology to explain the workings of autonomous systems more effectively, and also to produce autonomous systems which are easier to explain.

But the second difficulty is so far underexplored, and is a more important one for law and regulation. This is that the *kinds* of explanation required by law and regulation are not, at least at first sight, the kinds of explanation which AI systems can currently provide.

To answer the normative questions, law and regulation seeks a narrative explanation, a story. Humans usually explain their decisions and actions in narrative form (even if the work of psychologists and neuroscientists tells us that some of the explanations are devised *ex post,* and may not accurately reflect what went on in the human mind). At present, we seek these kinds of narrative explanation from AI technology, because as humans we seek to understand technology's working through constructing a story to explain it. Our cultural history makes this inevitable – authors like Asimov, writing narratives about future AI technologies like intelligent robots, have told

us that they act in ways explainable by the narrative logic which we use to explain human actions and so they can also be explained to us in those terms. This is, at least currently, not true.

This chapter argues that we can only solve this problem by working from both sides. Technologists will need to find ways to tell us stories which law and regulation can use. But law and regulation will also need to accept different kinds of narratives, which tell stories about fundamental legal and regulatory concepts like fairness and reasonableness that are different from those we are used to.

1 Introduction

Non-lawyers think that all law consists of rules, but lawyers know that much of it is a series of questions. This is particularly so when a legal system decides to regulate something, or when we are attempting to decide if some defect or failure should give rise to legal liability.

There are two main questions which we ask here:

- *What?* What ought to happen? What did happen? What should have happened?
- *Why?* Why will it happen? Why did it happen? Why wasn't it prevented?

These questions have served us very well when regulating human actions and deciding on liability where those actions cause loss or damage. But they work less well if we remove the humans from the loop[1] and instead hand over the decision-making and resulting actions to AI systems.

One reason for this difficulty is that these questions are primarily normative, not factual. The most important aspects of their answers, for law and regulation, tell us about how events *ought to* have occurred compared to how they actually did. When we ask them of the humans

1 An AI system which is 'human in the loop' makes a recommendation to a human, but the ultimate decision is still left to that human. The traditional questions asked by law and regulation can thus be applied to that human decision. The oft-expressed fear that humans will automatically assume that a computer's advice is more credible than their own judgment seems, according to empirical research, to be a myth – BJ Fogg & Hsiang Teng, 'The elements of computer credibility' (1999) CHI '99: Proceedings of the SIGCHI conference on Human Factors in Computing Systems May 1999, 80, 81.

who made decisions and initiated actions, we are trying to find out if those humans acted properly. We, or more accurately law and regulation, have over the years established standards for proper human behaviour. We know how humans ought to have behaved. But we are far less sure how AI systems ought to behave.

As an example, take the well-known Tesla crash in the US in 2016. It appears that an important cause of the crash was that the autonomous driving technology misidentified another vehicle as being part of the sky, and so did not brake or turn to avoid collision.[2] No human driver ought to make such a mistake, or rather, no human driver ought to make such a mistake for this reason. And yet, up until this crash, Tesla cars had driven themselves on the roads with far fewer accidents of any kind than would have been caused if they had been driven by humans. On one measure, the technology performs worse than humans; on a different measure, it performs much better. Which is the correct standard? Or is it neither?

For liability, this problem is one which time could solve. Through several hundreds of decisions about liability for crashes involving autonomous vehicles, the courts of each country would be likely to evolve suitable standards of performance for AI systems. Admittedly, we might not wish to live with the uncertainty until this evolution is complete, and there would still be uncertainty about how well newly developed AI systems met those standards, or whether the standards should later evolve to reflect improvements in AI design.

Regulators cannot wait that long. Their job is to devise regulations which mitigate the risks to society caused by the activities they regulate.[3] This requires them to set some standards in advance,

2 Larry Greenemeier, 'Driverless Cars Will Face Moral Dilemmas', *Scientific American* 23 June 2016, http://www.scientificamerican.com/article/driverless-cars-will-face-moral-dilemmas/; Tesla Motors statement, 30 June 2016 – https://www.teslamotors.com/en_GB/blog/tragic-loss.

3 In some cases, the mitigation might be through prohibiting the use of AI for a particular purpose – see, e.g., the list of prohibited AI practices in Article 5 of the proposed EU Artificial Intelligence Act (Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act), COM(2021) 206 final 21 April 2021). Explaining the decision-making of such an AI would not alter the prohibition, and therefore such AI systems fall outside the scope of this chapter.

That said, a decision-making explanation might be useful in deciding if the AI falls within a prohibition. For example, Art 5(1)(b) of the proposed AI Act prohibits use of

rather than waiting for the risks to eventuate and then deciding retrospectively what should have happened instead.

2 Explanations

In order to do their jobs, courts and regulators need answers to the *what* and *why* questions. In a world of human decision-makers these answers come in the form of explanations.

Let us suppose that a doctor misdiagnoses a patient's condition. On its own, this tells us nothing about whether the doctor failed to meet a normative obligation. Even if all the necessary standards are met, some medical diagnoses will be wrong. So instead, we interrogate the process through which the doctor made the diagnosis: what information did she take into account or ignore, and what were her thought processes when deciding what diagnosis to give based on the information she considered relevant? That explanation is given in a narrative form – it is the *story* of how the doctor undertook the diagnosis.

Now let us suppose that an AI system is undertaking the diagnosis. The obvious course of action is for a regulator or a court to demand a similar explanation, a story about the AI's decision-making processes. Such an explanation is the most important element of transparency, which has been recommended as one of the main tools for AI regulation.[4] The aim is that the AI, or its developers, should be able to explain the decision-making options available to the technology in each case, and the choices it made between them. If achievable, this would help resolve responsibility and liability

AI which exploits vulnerable persons, and whether or not such exploitation was occurring might not be knowable if the AI's decision-making cannot be explained.

4 See European Parliament resolution of 20 October 2020 with recommendations to the Commission on a framework of ethical aspects of artificial intelligence, robotics and related technologies (2020/2012(INL)), which calls for explainability in addition to transparency.

See also European Commission, *White Paper On Artificial Intelligence – A European approach to excellence and trust*, COM(2020) 65 final 19 February 2020; EU High-Level Expert Group on Artificial Intelligence, *Ethics Guidelines for Trustworthy AI*, 8 April 2019; UK House of Commons Science and Technology Committee (2016) *Robotics and artificial intelligence* (HC 145 12); US Department of Transportation/NHTSA (2016) *Federal Automated Vehicles Policy – Accelerating the Next Revolution in Road Safety*.

questions and assure regulators that the AI will not cause unexpected individual or social harm.

The question is whether the kind of narrative explanation that regulators and courts expect is actually achievable for AI.

2.1 Explanation through metrics

The easiest explanation which can be offered about an AI's decisions is given in numerical form, setting out how the AI has performed according to some chosen metric. Thus, the developers of a facial recognition AI might demonstrate that it can recognise faces it has previously 'seen' with 95% accuracy, or an autonomous vehicle might be shown to have 80% fewer accidents per 10,000 kilometres than human drivers do on average. This tells us something about how well the AI performs its task overall, but little or nothing about *how* it does so in each individual case. These metrics can also be misleading if the data are unrepresentative of the real-world cases in which the AI system could be used.

An additional issue with metrics is that there are multiple measures of performance which could be chosen when developing the AI. Optimising its performance against a particular metric may not optimise for the other metrics which could have been chosen. Our autonomous vehicle might have fewer accidents than human drivers, but more fatal accidents, and this might not be a better outcome overall. Therefore, an important facet of explainability lies in choosing an appropriate metric to evaluate the performance of an AI system.[5] And metrics are always a proxy for what we *really* want to assess; in this case whether the autonomous vehicle is safe enough for use on the road.

Further, there might be multiple AI solutions to a problem which score differently on the chosen metric, but one of those lower scoring solutions could still be preferable to the other choices. For example, a disaster response robot could choose a longer path to reach its objective as it avoids going through a weakened building that might collapse – something that is worse if measuring time to objective or fuel consumed, but is better when measured against the risk of damage to the robot and the likelihood of completing the mission.

5 Maria Fox, Derek Long, and Daniele Magazzeni. 'Explainable planning'. *arXiv preprint arXiv:1709.10256* (2017).

From a regulatory standpoint, the choice of metrics used when optimising and testing an AI is an important issue. But it should by now be clear that metrics alone are not enough to satisfy the explanatory demands of law and regulation. Something closer to how humans explain their actions will be needed.

2.2 Asimov explanations

It is worth repeating here the questions set out in section 1 which law and regulation ask about decision-making:

- *What?* What ought to happen? What did happen? What should have happened?
- *Why?* Why will it happen? Why did it happen? Why wasn't it prevented?

When humans are being regulated, we seek answers in the form of a narrative, explaining how the human went about making the decision in question. Then we can compare this answer to our chosen standard of human behaviour, such as taking reasonable care.

If we seek similar explanations about how an AI made its decisions, we are asking for what we, the authors, will call 'Asimov explanations'.

Stories of intelligent machines have been with us for millennia.[6] In *Politics*, Aristotle wrote:

> … if every instrument could accomplish its own work, obeying or anticipating the will of others, like the statues of Daedalus, or the tripods of Hephaestus … chief workmen would not want servants, nor masters slaves.[7]

Around a thousand years later, and about a thousand years ago, the Indian story book *Śṛṅgāramañjarīkathā* told of King Bhoja's pleas-

6 For a helpful overview of the earliest stories, see A History of Artificial Intelligence: Antiquity, https://ahistoryofai.com/antiquity/.

7 Aristotle, *Politics* (trans Benjamin Jowett, Oxford: Clarendon Press 1885) vol 1, 6; Book I part IV.

ure garden which contained a doll who could speak, along with a range of other automata.[8]

But the most influential stories about intelligent machines are undoubtedly those of Isaac Asimov, who published stories on this topic in the 1940s in the magazines *Super Science Stories* and *Astounding Science Fiction*, and then published them in book form as *I, Robot* in 1950.[9] In these stories, intelligent robots are constrained to obey the three laws of robotics[10] that Asimov invented. The stories explore the logical contradictions between these laws, which result in the robots behaving very differently from what was expected.

The importance of these stories is that the decisions and actions of the robots are explained to humans in terms of human logic. Observers of the robots induce their 'reasoning' and explain it using human language. These explanations are given as a narrative of the robots' 'thought' processes, and explain those processes just as a human actor might explain their own actions or decisions (or more accurately, as a human acting solely in accordance with a set of rules might do). Asimov's stories contain internal stories about how robots think, and they tell us that robot thinking can be explained via telling stories.

This cultural understanding that intelligent machines can be explained via stories has led to proposals to regulate AI by demanding narrative explanations about how it makes decisions[11], or even the imposition of express regulatory obligations to produce such explanations.[12] These demands, expressed through law and regula-

8 See Daud Ali, 'Bhoja's Mechanical Garden: Translating Wonder Across The Indian Ocean, Circa 800–1100 CE' (2016) 55 History of Religions 460, 462–3. The article later discusses other depictions of automata, many of which act autonomously, in Indian stories of that period.

9 Isaac Asimov, *I, Robot* (Gnome Press 1950).

10 1. A robot may not injure a human being or, through inaction, allow a human being to come to harm.
2. A robot must obey orders given it by human beings except where such orders would conflict with the First Law.
3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.

11 *Draft Report with recommendations to the Commission on Civil Law Rules on Robotics* (2015/2103(INL), European Parliament Committee on Legal Affairs 31 May 2016).

12 See e.g., *Federal Automated Vehicles Policy – Accelerating the Next Revolution in Road Safety* (US Department of Transportation/NHTSA, September 2016). Article

tion, are based on a belief that such explanations are possible. But our human beliefs about what is possible (apart from the beliefs of those who have studied AI technology closely) are culturally derived, originating in fictional narratives rather than scientific papers. They are likely to be wrong.

As we will see, AI cannot currently be explained in this way, and might never be able to explain itself solely by means of stories. This chapter therefore needs to investigate what kinds of explanations *can* be given.

3 The black box problem

Technical systems whose workings are not understandable by humans are often described as 'black box' systems. Some AI systems are not black boxes in this sense – for ones that use simpler mechanisms, it is possible to accurately describe the processes through which the AI reached its decision. Such a system is inherently interpretable, and an interpretation of a decision is a full, though highly technical, explanation of how that decision was arrived at.

But from the perspective of law and regulation, a technical interpretation might be equally as opaque as a true black box system. The relevant question, from that perspective, is whether the person who is entitled to ask the question can understand the explanation. If not, the AI is functionally a black box in this context, even if in some other context (AI development, for example) the explanation might be comprehensible. For example, the developer of a machine learning-based AI might be able to explain to another AI developer how and why the AI reaches its decisions, but that explanation tells the user of the AI nothing. All that the user knows is that he is ignorant of the AI's workings, and that it is *de facto* a 'black box'.

13(2)(f) of the EU General Data Protection Regulation, Regulation 2016/679, entitles data subjects to 'meaningful information about the logic involved' in automated decision-making involving their personal data. The proposed EU Artificial Intelligence Act (n 3) adopts a more nuanced approach in article 13(1): 'High-risk AI systems shall be designed and developed in such a way to ensure that their operation is sufficiently transparent to enable users to interpret the system's output and use it appropriately.' See also Article 14(4)(c) requiring that those responsible for high-risk AI systems should be able to correctly interpret their outputs.

In this sense, the opacity of the AI system also depends on when and how the question is asked. If the AI has produced a result which causes loss or damage, it may be possible to obtain some kind of answer depending on what type of AI system it is. However, explanation in advance, to help a regulator decide if an AI meets any requirements necessary for its use, is more difficult. In terms of their capacity to have their decision-making explained, AIs can be classified into two types.

Rule-based AI technologies implement sets of rules (analogous to IF … THEN … statements), and these sets of rules result in a decision tree. In theory, these rules could be hand-crafted, and the person doing so could therefore explain the decision-making process in terms which a human might understand. Each decision by the AI is the result of a single path through the decision tree to the output, and that path could be described as the 'reasoning' which led to its decision. However, all but the simplest rule-based AIs are likely to generate their rule sets through machine learning processes, such as genetic techniques which combine parts of two current rule sets and keep the 'offspring' which perform better than their parents. The resulting rule set is thus not an implementation of the reasoning processes of a human mind. If it were subsequently analysed by a human, some description of its reasoning for an individual decision could be produced, but that description will be of complex and technological reasoning, and unlikely to produce the kind of narrative explanation that non-technologists understand. Stories about human decision-making concentrate on motivation and intention, neither of which will be present here. There is also a likelihood that the logic of the resulting rule set may well be too different, detailed and complicated for the human mind to understand fully, what Burrell describes as:

> opacity that stems from the mismatch between mathematical optimization in high-dimensionality characteristic of machine learning and the demands of human-scale reasoning and styles of semantic interpretation.[13]

13 Jenna Burrell, 'How the machine "thinks": Understanding opacity in machine learning algorithms' (2016) Big Data and Society 1, 2.

Pattern-matching AI technologies such as neural networks do not make decisions by following a path through a decision tree. They identify and match patterns in their inputs, and from those patterns they induce (rather than deduce) their output.[14] These systems are highly probabilistic – the output of an image recognition AI would not be 'this picture is of a moose' but rather 'this picture is more likely of a moose than any other animal' (possibly with a probability value for that likelihood). The AI learns how to make its decisions by analysing a large and comprehensive training dataset, and is then tested against a substantial real-world dataset. This process is iterated until the AI succeeds on real-world data sufficiently well to be put into use. From a non-technologist perspective, it 'just knows'. This makes it difficult to explain how the technology came to its decision, and thus how any loss or damage was caused. It is likely to be near-impossible to explain it in narrative terms.[15] Even if a rule set approximating the AI's decision-making could be reverse engineered, those rules might not convey anything meaningful to humans – 'IF pixel at address X,Y has colour value > N THEN …'.

For both technologies, after-the-event explanations are often possible, although they may only be properly comprehensible to a few, highly-qualified humans. What, though, of explanations in advance, before the AI system is put to use? Regulators, and others such as insurers, might well want such explanations to assess the risks which arise from using the AI and how well they have been anticipated and guarded against. And the wider public might want such an explanation to persuade them to accept the technology – most citizens would be unconvinced by autonomous vehicles if all that they were told was, 'We can't explain how it works, but it's really safe.'

Generating an explanation in advance through human analysis of an AI's workings is particularly difficult. Algorithmic AIs are hard to explain because there are so many paths through the decision tree, maybe millions of paths in some cases. Small changes in inputs can result in very different outcomes. Explaining all these paths will not

14 In some cases, the human developer instructs the system what it should be looking for (supervised learning), in others the system just learns whatever it can (unsupervised learning).

15 Burrell, n 13, 5–7.

provide what is wanted – the human need for narrative requires an abstraction, a coherent collective story into which all these different paths fit. Such a narrative might not even exist; if it does, the human mind may not be up to the task of constructing it. For example, devising an advance narrative explanation of the workings of a neural network is a particularly intractable problem for human analysts, because there is no logic (in the human sense) behind its decisions.[16]

All this suggests that human creators of AI will rarely be able to provide the narrative explanations which law and regulation currently demand. This is a problem, because demanding narratives as a precondition for allowing use of an AI (or granting insurance, which is a precondition of use if the AI producer wishes to avoid insolvency) will in many instances amount to prohibition on using that AI at all.

4 Technology tools for explanation

So can technology help us to produce the explanations we want for law and regulation? There are two parts to this question. The first is what technology can actually tell us about the decision-making processes of AIs, both in advance and after the event. The second is how we can fit that information into our legal and regulatory explanation-demanding systems.

Answers to the first part are likely to come from the fast-developing field of eXplainable AI (XAI). The goal of XAI is to design tools that can provide explanations for the decisions of complex autonomous systems. The purpose of these explanations is to assist humans to understand the decision-making process, focusing on a number of key drivers. These include confidence, trust, safety, ethics and

16 Humans are happy with making illogical decisions, of course. The music or food one likes is not decided through logical processes. But these kinds of decisions are deliberately excluded from the sphere of law and regulation. Where an activity falls within the legal and regulatory sphere, humans are expected to give narrative and logical explanations of their actions. The explanatory logic used in law and regulation tends to be simple propositional logic, for example: 'IF it is snowing THEN drive slower'.

fairness.[17] By exposing the reasoning of an AI system, XAI can lead to improved performance in future iterations.[18]

4.1 XAI techniques

Developments in XAI are advancing rapidly, and there is as yet no consistent terminology or taxonomy of XAI techniques. However, a recent survey of the field[19] suggests that the following categories of XAI research might usefully group related techniques together:

1. Saliency techniques. These identify the relative importance of different inputs to the AI in producing particular outputs – for example, the regions of tissue that contain cancerous cells. Results are often presented visually or quasi-visually (e.g., in the form of a heat map of words or phrases for textual analysis AIs). The idea here is that these representations will produce patterns which humans can map to their own understandings of how decisions in that field are made, and thus use them to explain the AI's decision-making.
2. Signal methods. These are used for image recognition neural networks, and identify how input images affect the values of the neurons in a layer of the network. What that layer 'sees' can then be reconstructed and compared by a human to the original image, to discover which parts of the input image are detected by each layer. From this a narrative might be constructed in the case of, say, facial recognition: 'First the AI identifies the eyes and nose, the next layer finds the edges of the face, the third layer …'.

17 Doran, Derek, Sarah Schulz, and Tarek R. Besold. 'What does explainable AI really mean? A new conceptualization of perspectives'. *arXiv preprint arXiv:1710.00794* (2017).

18 Anjomshoae, Sule, et al. 'Explainable agents and robots: Results from a systematic literature review'. *18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), Montreal, Canada, May 13–17, 2019*. International Foundation for Autonomous Agents and Multiagent Systems, 2019.

19 Tjoa, Erico, and Cuntai Guan. 'A survey on explainable artificial intelligence (XAI): towards medical XAI'. *arXiv preprint arXiv:1907.07374* (2019). For alternative taxonomies, see e.g., Biran, Or, and Courtenay Cotton. 'Explanation and justification in machine learning: A survey'. *IJCAI-17 workshop on explainable AI (XAI)*. Vol. 8. No. 1. 2017; Guidotti, Riccardo, et al. 'A survey of methods for explaining black box models'. *ACM computing surveys (CSUR)* 51.5 (2018): 1–42.

3. Verbal (or textual) interpretability methods attempt to translate symbolic processing into verbal 'IF … THEN …' rules. These methods are likely to be used on text analysis algorithms, because the input text can be used to construct the 'IF … THEN …' statements which explain the AI's decisions. In effect, these statements are a higher-level abstraction of the more complex set of rules actually embedded in the algorithm. One known problem with verbal interpretability is justifying the techniques used to produce the verbal 'IF … THEN …' statements – these techniques might still be 'black boxes' so far as the person receiving the explanation is concerned.
4. Mathematical modelling. This technique requires a mathematical model to be devised which matches (or perhaps more accurately: approximates) the relationship between inputs to the AI and its outputs. A technical expert will be able to understand that model, and it is hoped will also be able to explain it in non-mathematical terms to any human who requires an explanation. In effect, the human-incomprehensible workings of the AI are abstracted into a mathematical model which is understandable by some skilled humans, and those humans can explain them to other humans at an even higher level of abstraction.
5. Feature extraction (or importance). This identifies features in the input data (e.g., for medical diagnosis, the inputs relating to fitness, eating patterns and sleep patterns) and then identifies the features which are most strongly correlated for particular outputs and those which are not correlated. Feature extraction is thus a type of abstraction; it might find, for a particular disease, that when the AI makes its diagnoses, sleep and diet are closely correlated, whereas geographical residence and income are not. These correlations can be used for human explanations.
6. Sensitivity methods. These take individual decisions of the AI and make changes to its inputs, to see how they affect its outputs.[20]

20 There is a growing legal literature on counterfactuals, which are a type of sensitivity method. Counterfactual explanations function by reiterating a data process with the smallest possible change to determine which parts of the data are influencing a decision. Small tweaks are made to the data, then the 'question' put to the AI is asked again and again, pinpointing which data points changed the outcome.

'In the existing literature, "explanation" typically refers to an attempt to convey the internal state or logic of an algorithm. In contrast, counterfactuals describe a dependency on the external facts that led to that decision.' Sandra Wachter, Brent Mittelstadt

This can identify which inputs are most important for producing the decision. It can also offer a measure of reliability for the AI, because if tiny changes in inputs produce major changes in output, the AI might not produce reliable results on inputs it has not seen before. One difficulty with these techniques is generalising them to provide useful information about the workings of the AI overall, rather than just explaining individual decisions.

7. Optimisation (or decomposition). This attempts to find sub-elements of the AI which, for the same input data, produce outputs which are recognisably related to the full AI's output. This is a kind of abstraction of the AI, and the theory is that the abstraction can be interpreted (probably by technical experts) to discover information about the full AI's decision-making.

From these descriptions it is clear that there is no single tool which will be able to provide the explanation needed by law and regulation.[21] Different explanations are needed by different users of explanations, for example the explainability requirements for a regulator or a developer would be different to those needed for an end user.[22] However, each tool potentially contributes something useful, and they might be used in combination to assist the explanation process.[23]

---

and Chris Russell, Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR (2018) 31 Harvard Journal of Law & Technology 841, 845.

For a full discussion of counterfactual explanations, see Katja de Vries, Transparent Dreams (Are Made of This): Counterfactuals as Transparency Tools in ADM (2021) 8 Critical Analysis of Law 121.

See also Atoosa Kasirzadeh and Andrew Smart, The Use and Misuse of Counterfactuals in Ethical Machine Learning [2021] arXiv:2102.05085 [cs] <http://arxiv.org/abs/2102.05085> accessed 24 February 2021.

21 Indeed, some tools are developed specifically to explain a particular AI's decisions, and thus would not be usable to explain other AIs.

22 Sam Hepenstal and David McNeish. 'Explainable Artificial Intelligence: What Do You Need to Know?'. *International Conference on Human-Computer Interaction*. Springer, Cham, 2020.

23 Langley, Pat, et al. 'Explainable agency for intelligent autonomous systems'. *Twenty-Ninth IAAI Conference*. 2017.

4.2 Using the XAI tools to explain

How these tools might be used depends very much on how well two factors are understood:

- The input data which the AI might receive; and
- The consequences for the external world which that AI's outputs might have. The range of decisions it can make will of course be known, but the potential consequences of those decisions might or might not be known, or even knowable.

4.2.1 Inputs

For domains that have well-understood inputs, it is possible to have an understanding of how the system *should* work. This means that any explanations generated for an AI working in such a domain should match the expectations of humans who currently work in the domain. For example, in medical imaging the range of images which might be assessed is known, and doctors already know what they are looking for in those images. Thus, if they are provided with the explanations from an AI system, they can verify that the rules or techniques which the AI appears to have learnt match the image analysis rules which they apply themselves. An XAI explanation which highlighted the elements of an X-ray that leads the AI to a positive classification of cancer, for example, could be used by doctors to check whether these are the same elements which guide their own diagnoses.

However, in other domains we might not have a good understanding of the inputs, or the input space might be so large that the AI cannot be trained on every possible input it is likely to encounter. An autonomous vehicle used on Canadian roads might expect to encounter a moose or a bear, and thus be trained to recognise those animals, but a peacock would be as much of a surprise to that vehicle is it was to one of the authors when he encountered one on an English country lane.

If some of its inputs are unknowable in advance, it is hard to say how an AI *should* work. Even if we can explain how it will behave if it encounters a moose or a bear, we can only guess what it will do when presented with a peacock. This does not mean, though, that XAI cannot provide some assistance, particularly in open-ended domains where the optimal strategy is unknown to humans. As an

example, we can be left scratching our heads when an AI system outperforms us and we don't know how it makes its decisions. DeepMind's AlphaZero made radical and unexpected moves in the game of Go which ultimately proved to be beneficial later in the game, and expert players are still studying and analysing those moves.[24]

The explanations which XAI can provide can help in two ways here. First, they can expose some information about the AI's decision-making and thus provide some reassurance that it is not doing something untoward, such as making unlawfully biased decisions.[25] Second, they can increase our knowledge about a domain (e.g., by highlighting a previously unknown relationship or explaining why a particular course of action is beneficial). From a regulatory perspective this is helpful in ensuring the system aligns with long-term goals, such as improving industry standards, by revealing something new about how good performance can be achieved.

4.2.2 Consequences

When it is foreseeable that the outputs of a system might produce consequences which society will wish to avoid, such as deaths on the roads or inaccurate medical diagnoses, regulation attempts to ensure that these foreseeable failures do not occur. This entails putting the system in scenarios where a foreseeable fault could occur, and testing to see if it still acts as intended. XAI could assist in testing AIs by going beyond just observing the system's behaviour; it might allow the developers to ensure that the AI actually recognises the potential failure and takes steps to avoid it, rather than simply succeeding by some fluke occurrence. An example would be exposing an AI to

24 Silver, David, et al. 'Mastering the game of Go without human knowledge'. *Nature* 550.7676 (2017): 354–9.

25 Let us imagine an AI which selects students for a drama degree. Anecdotally, the culture of the acting profession has been welcoming towards those of a minority sexual orientation, which might attract such persons to attempt to enter the profession. Our AI, learning from previous applications and examples of accepted students, might therefore teach itself to rely on clues to sexual orientation in deciding which students to select. This would be unlawful, so an explanation sufficient to show it is unlikely to be doing this would be useful, even if that explanation cannot give a full picture of how the AI works. For a non-fictional example of unintended bias derived from AI training data, see Jeffrey Dastin, Amazon scraps secret AI recruiting tool that showed bias against women, Reuters 11 October 2018 (https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G).

adversarial examples designed to catch the system out, and seeing if it fails. If so, XAI tools will help in explaining why it fails so that developers understand how the system can be modified to avoid that failure in the future.

Unforeseeable consequences must be expected when it is impossible to test the AI system in every possible scenario it will ever encounter. The regulatory problem here is achieving sufficient reassurance that the AI will (or is at least likely to) act correctly in these circumstances, because the potential consequences of its decisions will by definition be unforeseeable.

If the internal decision-making of the system can to some extent be understood through explainability tools, and it is believed to fit well enough with known human decision-making in the same domain, a regulator might treat this as adequate assurance that if the AI encounters previously unseen situations it is likely to produce decisions whose effects are unlikely to be harmful (or at least, no more harmful than the effects of a human decision in such cases). XAI might even provide greater assurance than is achievable for human decision-making here, by finding the edge cases where the AI would act unexpectedly, exposing where there is a risk of unforeseeable consequences and perhaps even enabling the likely effects of the decision to be predicted. Requiring developers to use XAI tools to achieve a better understanding of how an AI makes decisions and how those decisions will affect the world around it might be a useful regulatory intervention for some types of AI.

5 Reconciling explanations

So what does this tell us about what XAI can offer, working in conjunction with human AI developers, to explain AI to law and regulation? We could bring this all together, in broad terms, as follows.

a) The easiest explanation which can be given for an AI is some suitable metric about its performance. This might compare the AI's decision-making numerically to that of a human undertaking the same task, or it might explain what proportion of the cases it was tested on were decided correctly, as assessed by its human developers. These numbers are useful as one factor in deciding if the AI is sufficiently good at its task to grant it regulatory approval, if

approval is needed, or to help decide if those producing or using it were in breach of their legal duties if a liability claim is made. However, the numbers only tell us about the overall performance of the AI – they give no clue about how well it decided in any individual case, or how it will perform in future cases.

b) XAI is sometimes able, in advance of an AI being put to use, to generate some information about the robustness and accuracy of the AI's decision-making. This will be by categorising some factors or reasons which are common to cases where the AI failed to make the decision which a human should have made, or which humans assess that the AI should have made. This might similarly be useful for deciding on regulatory approval or liability.

c) It can also be possible, in some instances, for XAI to identify which inputs most strongly influence the final decision and which have little effect. In advance of the AI making a decision, this will be an indication of which inputs are likely to be used in making a decision. After the event, it should be possible to say which inputs were or were not influential, though perhaps in terms of probabilities rather than certainties.

d) XAI might also be able to explain, to some extent, the order in which an AI builds up its decision, which could tell us something about dependencies. In the facial recognition example above, it might reveal that accurate identification of facial shape depends on accurate identification of eyes and nose, and so on. This could form the basis of a narrative explanation about how the AI is, or more accurately might be, working.

e) In the best case, from a legal and regulatory perspective, XAI might even produce an abstracted, high-level explanation of the *likely* 'reasoning' which a particular AI is using. But law and regulation will need to understand that this abstraction is a model of what the AI might be doing, which is developed from sample cases and the result of human interpretation of an XAI analysis of the AI's workings. The model might work only for some types of case, and not for others, so this is a dynamic explanation – over time the model might be disproved and an alternative model developed based on the XAI analysis, or XAI might improve the model so that its explanation is reasonably accurate for more cases. Because the model is both an abstraction and a simplification, it will not capture the full complexity of the decision-mak-

> ing, and thus cannot be relied on as a comprehensive explanation. Such a model is only the best guess that can currently be achieved about the AI's 'reasoning', a mixed product of machine analysis and human interpretation.

Working through a hypothetical example of an after-the-event explanation might be useful. Suppose that a fully autonomous vehicle collides with a pedestrian who has stepped into the road. What could an explanation aided by XAI look like compared to the explanation of a human driver?

A metrical explanation, or one which focuses on the general reliability and robustness of the AI (points a and b above), is of little help here. These are only useful in explaining whether it was safe to use the vehicle on the road at all, and we can assume that the fact that it was permitted on the road by regulators and insurers means that it had passed that test. So we might expect an explanation something like this:[26]

- Factual data from sensors tells us about the speed the vehicle was driving, light conditions, etc.
- The AI driving technology identified that there was something in the road, but initially misidentified it as most likely (probability 0.82) a black plastic bag blowing in the wind and so did not slow down.
  - This happened because lidar[27] signals, used to identify 2D outline, colour etc, are processed faster than sonar signals, which contain supplementary information about the 3D shape of objects.
  - The obstacle recognition element of the AI identifies outline first, in this case as being probably that of a plastic bag.

26 This hypothetical explanation is loosely based on the Uber autonomous vehicle crash in Arizona, March 2018. See NTSB Report NTSB/HAR19/03, https://www.ntsb.gov/investigations/AccidentReports/Reports/HAR1903.pdf; NTSB Board meeting documents, 19 November 2018, https://www.ntsb.gov/news/events/Pages/2019-HWY18MH010-BMG.aspx.

27 A technology commonly used for autonomous vehicles which uses the return signals from lasers to calculate distance from a target (here, the pedestrian) and also to create a 3D representation of the target.

- The AI then identified the obstacle as probably being a person (probability 0.91) using the additional data and braked, but there was insufficient time to stop before the collision.
  - As further data comes in, the nature of the obstacle is recalculated, adding revised lidar and sonar data as available.
  - Sonar data is more influential than lidar data in making the decision to brake (probability(sonar)*0.6 + probability(lidar)*0.4).
  - The model of the AI's reasoning suggests that assessment of the obstacle as more likely human than plastic bag using lidar data, and receipt (but not processing) of sonar data which would indicate its shape fitted human better than plastic bag, both happened at about the same time.
  - Braking started almost immediately thereafter (0.27 seconds).

The explanation of a human driver would be much briefer, something like this:

- I was driving below the speed limit and the light was poor, so I was keeping a good lookout.
- I saw the pedestrian in the road, but thought he was a black plastic rubbish bag blowing in the wind because his dark coat was flapping, so I didn't slow down immediately. In the circumstances, another human driver would have made the same misidentification.
- When I realised it was a pedestrian I braked hard, but this was too late to avoid the collision.

The first thing to note about these two explanations are that their main elements are broadly the same. However, for the AI, there is much more information about how it reached the various decisions it made.

The second thing is that the various explanations which XAI can provide about the AI driver do not form a coherent narrative about its 'motives' or 'intentions', which are an important part of the human driver's narrative. A human interpreter can take these XAI sub-explanations and weave them together to create something which approximates to such a narrative, but this is not the AI explaining itself – it is a human, generating an Asimov explanation

of the AI's decision, based on observation of its workings by XAI tools.

The third thing is that the explanation given about the AI is largely probabilistic, except for the data about speed and light, which are objective. By contrast, the human driver's explanation is set out in definitive terms and is deterministic. Further thought should tell us, though, that the human driver's explanation is less reliable than that given by the AI. It depends on how accurately the driver can recall her speed, the thoughts which were going through her mind, and so on.

At first sight, these two explanations seem quite different. The explanation of the AI tells us what probably happened in decision-making, with reliable data to support those probabilities. The human explanation tells us definitively what is claimed to have happened, but without reliable data to support it.

Further thought should tell us that in fact the human explanation is also probabilistic. We cannot be sure that it is correct, and so for legal purposes we have to make an assessment about how probable it is to be accurate. In a civil action, for example, we would ask whether the driver's version is more likely than not (probability 0.51 or greater) to be a true recollection.[28] Both explanations, AI and human, are uncertain. We might even argue that the main difference between them is that the AI explanation admits the uncertainty.

What, though, if we are asking a similar question in advance of there being an accident? That question might be whether the human driver, or the autonomous vehicle, will drive safely enough to be permitted onto the road at all.

For autonomous vehicles, this is where metrics and assurances about the reliability and robustness of the AI will come into play. Some measure of safety can be derived from comparative accident statistics about this AI's driving compared to that of human drivers, and is likely to be highly favourable to the AI or it would not be a commercially viable proposition. If a regulator wanted greater reassurance about particular driving situations where doubts had been

28 Noting also, of course, the extensive body of psychological research which indicates that human memory can be distorted by belief. Thus, a driver who believes that he is a safe driver is likely, without intending to do so, to revise his memory of an accident to fit in with that belief. See further Rodriguez DN & Strange D, 'False memories for dissonance inducing events', (2015) 23(2) Memory 203.

raised, this might be provided by reviewing training failures as if they were real accidents and seeking the kinds of explanation set out above. If a generalised model of the AI's reasoning could be produced by XAI, the regulator could compare that to how humans are believed to make driving decisions in order to identify differences or gaps. Lastly, the AI developer's plans and processes to monitor and improve performance, particularly through analysing accidents, will be an important factor in deciding whether sufficient safety is likely.

Human drivers have it much easier. The majority of safety assurance is achieved through the training and examination required for a driving licence, and after that drivers are incentivised to continue to drive safely by criminal sanctions and legal liability, reflected in insurance premiums.

In both cases, the answer to the question is in fact a prediction, that the human or the AI will drive safely. For the human, that prediction is based on passing a driving test and the hope that the legal and financial incentives to drive safely will be effective. For the AI, there is likely to be more evidence on which to make the prediction, but as humans we find it hard to evaluate whether this evidence is more or less reliable or objective than the evidence underpinning our prediction for the human driver. Members of society, and regulators, are humans, and thus have an intuitive understanding about the reliability of predictions about other humans. AI reliability cannot be evaluated in the same way.

6 Conclusion

As we have attempted to show in the previous section, a detailed analysis of the explanations which humans give for their decisions, and those which XAI might enable to be given about an AI, shows that they are likely to be much closer to each other than appears on the surface.[29] And yet our first instinct as lawyers and regulators is to accept the human explanations but reject the AI explanations as inadequate. Why might this be so?

29 Though we should note that this conclusion may not hold for all domains. As a simple example, an AI controlling a home heating system will be making very different decisions from a human controlling the same system manually, though their end aim (a comfortably warm home) is of course the same.

Clearly it is the fault of Asimov and other tellers of fictional tales about intelligent machines. Humans explain themselves in definitive terms – this is how it happened, this is how I will decide – but AIs are predicated on uncertainty and only tell us probabilities – this is most likely to be how it happened, this is probably how future decisions will be made. XAI-assisted explanations for what has already happened might, as we have seen above, be similar enough to human explanations once their probabilistic nature is understood. Explanations about the future decisions an AI might make are, though, very different from those about human decision-making.

This clash of narrative expectations seems a plausible reason why we might demand more from an AI by way of explanation. But an AI whose future actions can readily be explained in deterministic terms – what *will* happen in its 'reasoning', not what is *probable* to happen – is likely to be much less able, and thus less useful, than the kinds of AI we have been discussing in this paper.

If we wish to secure the likely benefits from those kinds of AI, we will need to change our attitude to explanations. After all, the certainty which human explanations appear to offer is, we suggest, a false certainty. If we can accept that explanations for highly complex systems (including humans, who are highly complex) must inevitably be based on probabilities, we will have made a useful advance in law and regulation.

How to Regulate AI?*

PETER WAHLGREN

> Again and again I have heard the statement that learning machines cannot subject us to any new dangers, because we can turn them off when we feel like it. But can we? To turn a machine off effectively, we must be in possession of information as to whether the danger point has come. The mere fact that we have made the machine does not guarantee that we shall have the proper information to do this. This is already implicit in the statement that the checker-playing machine can defeat the man who has programmed it, and this after a very limited time of working in. Moreover, the very speed of operation of modern digital machines stands in the way of our ability to perceive and think through the indications of danger.
>
> *Norbert Weiner 1961*

Introduction

The question on how to regulate AI has recently attracted great interest and scholarly attention. For anyone with a background in the field of IT and law[1] this is inspiring but also somewhat surprising for several reasons. First, as illustrated in the citation above, the

* This article originates from the research project *Legislative Techniqu*es, financed by Torsten and Ragnar Söderberg's Chair in Legal Science.

1 For this author (a semi-autonomous human), this is a renewed address. Previous efforts related to methodological and regulative issues of AI were carried out between 1983 (*Artificiell intelligens, AI,* The Swedish Law and Informatics Research Institute, Stockholm 1983 (IRI:PM 1983:9)) and 1994 (Wahlgren, P., *A General Theory of Artificial Intelligence and Law*, in Legal Knowledge Based Systems: The Relation with Legal Theory. Prakken, H., Muntjewerff, A.J., Soeteman, A., Winkels, R.G.F. (eds) Koninklijke Vermande BV, (1994)), the major contribution being Wahlgren, P. Auto-

regulative problems originating from AI-related applications were identified several decades ago, and, accordingly, have been addressed extensively. Second, many recent arguments regarding ways of regulating AI reflect a shift from traditional legal mechanisms towards a focus on vaguer and less defined concepts lacking instrumental abilities and enforceable sanctions, such as ethics and morals.[2]

The overarching questions for a renewed investigation of regulative problems related to AI are thereby given. What is new, and if the aforementioned impression is true – i.e., that law is no longer seen as a first-hand remedy for dealing with upcoming problems – what are the reasons for this downgrading and what are the alternatives?

An analysis of the relations between law and AI from such a starting point presupposes a return to, and perhaps a reinterpretation of, several fundamental preconditions. Consequently, this article is a revision of a number of basic questions: what is AI, what is law, what is there to regulate and what kind of regulative tools are available?

1 What is AI?

Opinions about AI differ: bright, sceptic and dystopian outlooks and heavily clashing arguments about the future of the technology abound. This is not a new phenomenon. From the outset of its development, AI has generated criticism and the idea of intelligent machines has sometimes been ridiculed. In early development projects, especially in the legal sector and public administration, this led to strategic shifts in terminology. In order not to offend potential end users, hinting that they would soon be replaced by machines,

---

mation of Legal Reasoning: A Study on Artificial Intelligence and Law, Kluwer, Deventer (1992).

2 See, e.g., European Commission, High-Level Expert Group on AI, *Ethics guidelines for trustworthy AI*, https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai (2019); Government of Canada, *Responsible use of artificial intelligence,* https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai.html#toc1; Müller, Vincent C., Ethics of Artificial Intelligence and Robotics, Stanford Encyclopedia of Philosophy (2020), https://plato.stanford.edu/entries/ethics-ai/; Montreal AI Ethics Institute, https://montrealethics.ai/; *AI Ethics Guidelines Global Inventory, AlgorithmWatch's inventory of principles, voluntary commitments and frameworks for an ethical use of algorithms and AI*, https://algorithmwatch.org/en/ai-ethics-guidelines-global-inventory/ (2019).

or, if they held positions as independent decisionmakers, should soon be dethroned to keyboard typing clerks, AI as a concept was often consciously avoided. Intended applications were described as *decision support systems* made possible by the employment of methods described as *knowledge-based* or with in similar, less provocative terms.[3] The focus was not on the development of autonomous devices and the argument in favor of the technology was that it had the potential to enhance quality and efficiency. Although much of this is history now, it is obvious that manifold views still exist and it is easy to become bewildered. AI is a complicated subject matter and identifying essential components, functions and consequences is not always an easy task. A few personal observations might help to explain this state of the matter and, perhaps, provide a basis for a better understanding.

1. The ability to build intelligent machines is dependent on knowledge from many fields and AI research is in principle borderless. Methods and approaches vary significantly. Digital techniques and computers are core elements, but not only technical matters are relevant. In order to develop practical AI, input from a large variety of natural, human and legal sciences must be acknowledged.

From this follows that the underlying perspectives vary to a large extent – AI projects are initiated from various standpoints and with different objectives. The software programmer developing algorithms for emotional recognition has a different focus than the biochemist utilizing AI in order to identify the modus operandi of new virus mutations. The design of sensor-based systems for large-scale disaster management has little in common with 3D simulation of heart surgery for educational purposes, and there are countless other examples. An additional consequence of the broadness of the topic is that many projects are framed by different administrative frameworks, financing routines and traditional academic disciplines. The approaches therefore vary greatly and unintentional or unnoticed silo mentality may sometimes explain what appear to be varying opinions about how to prioritize and allocate resources.

Different understandings and expectations also appear in the meeting between AI and law – being a lawyer in AI development projects engaging large numbers of technical experts can be challenging. Explaining the

3 See e.g. Wahlgren, P, *Beslutstöd för brottmål* (*Decision support for criminal court cases*), Swedish National Courts Administration, Report 1987:9 and Wahlgren, P., *Swedish Experiences with Decision Support Systems*. In Expert Systems in Law, Proceedings from an International Conference on Law & Artificial Intelligence. Bologna May 3–5 1989.

necessity of implementing legal principles and security frameworks in complex systems can be problematic, especially if the requirements put forward presuppose extra design efforts or may be interpreted as something that hampers the efficiency of the systems. Although confusion and misunderstandings are usually possible to sort out, it is important to remember that *AI is an open, multifaceted discipline*, involving a large number of stakeholders with different backgrounds and objectives.

2. AI takes on different shapes. Robots and tangible AI devices perform myriad tasks in many sectors of society, e.g., in production, transport, health, logistics, and research. Many manifestations of AI are easy to spot and understand. In autonomous vehicles, for example, AI is an embedded sub-component. Most of its operative functions are nevertheless clearly visible and although long-term effects of machine learning may linger, consequences of functional errors or unacceptable deviations are likely to be observed immediately.

Components of AI can however also be intangible and exist only as algorithms cast into computer code. The functions may thus be concealed from the outside and integrated as tacit sub-processes in systems of systems, working with various degrees of autonomy at different levels. A self-learning ability may also lead to functional creep that is difficult to detect. Changes may be abrupt and unpredicted – obviously system component malfunctions could have this result, but also if unexpected threshold values are reached as data are accumulated, combined and processed.

AI is therefore not a well-defined wonder and attributes and functions are not always possible to identify. On the contrary, AI applications are uncountable and so are its purposes, methodological approaches and consequences, existing and potential. Some devices may perform the intended tasks with a high level of accuracy, others may generate unwanted or even unnoticed side effects, while still others can be criticized as biased or unethical, and so forth. Consequently, the manifestations of AI vary greatly and in its operative modes *AI performs innumerable tasks in significantly diverse applications*.

3. AI is seldom described in a uniform manner. The underlying focus can be any of many detailed technical levels, applications, functions or consequences. As the previous experiences, intentions and knowledge of the stakeholders differ, so do accounts and the ways in which similar things are presented. AI systems are frequently described as "autonomous" systems, but established definitions of autonomy exist in only a few domains. An additional important factor is that the terminology is often of a technical nature, difficult for the non-expert to grasp.

Over time, similar things are given different names and sometimes words have dual meanings. Words like "rule," "document," "file," "equal" and "fair" carry different meanings for lawyers, computer scientists and statisticians, and seemingly obscure conceptions abound. Conceptual confusion is a recurring problem when AI is to be integrated into various settings, each of which may have its own terminology and jargon. This is also a common complication in the meeting between AI and law. Deep learning, predictive modelling, autonomous algorithms, hard AI, black boxes, deontic logic, data subjects, privacy by design and teleological interpretation are just a few examples. Difficulties of this kind are possible to overcome, but it takes time to acquire the necessary domain proficiency, and becoming an AI generalist is a demanding task. Often there is a need for intermediaries that are able to bridge the knowledge gap and identify misconceptions. Thus, for the new arrival, *AI may stand out as a complicated and confusing subject matter.*

4. Dynamics is another crucial component of AI, partly because the concept of *intelligence* fluctuates. Although it is commonly accepted that intelligence is a generic ability depending on the activation of a large number of identifiable faculties, this interpretation has been contested. AI theorists have suggested that intelligence is merely a name for processes which are poorly understood and yet not programmed.[4] As our understanding increases and development progresses, the concept of intelligence is changing its meaning.

A shifting understanding of AI is also reflected in different generations of AI research. In the formative period, AI was by necessity a field of theoretical studies due to lack of technical means. This was followed by a period of experimental pilot system developments, eventually leading up to where we are today, with practical applications in daily use. From a methodological point of view there have been parallel shifts, from early efforts focusing on construction of relations in databases, followed by a focus on the elicitation, interpretation and transformation of knowledge into logical programs, towards an increasing interest in statistics and the development

4 *See*, e.g., Hofstadter, Douglas R., Gödel Escher, Bach: An Eternal Golden Braid, Penguin Books Ltd. (1979) p. 621: "[O]nce some mental function is programmed, people soon cease to consider it as an essential ingredient of 'real thinking'. The ineluctable core of intelligence is always in the next thing which hasn't yet been programmed" *and* Minsky, Marvin, The Society of Mind, Simon & Schuster Inc., (1985) p. 71, "Our minds contain processes that enable us to solve problems we consider difficult. 'Intelligence' is our name for whichever of those processes we don't yet understand."

of so-called machine learning methods, able to detect patterns in large datasets and adjust algorithms dynamically.

Inherent in the dynamics is that AI is a moving target, constantly changing with the evolution of the field. Blunt illustrations of this can be seen in the many functions and apps resulting from AI research and currently available in the mobile devices used every day. A couple of decades ago several of these facilities would have been dismissed as science fiction. Today, more than a few of them are seen as basic trivialities, if they are observed at all. Another example is the successive generations of telecommunications. Consequently, *AI is what comes next*, and this is sometimes fertile soil for speculations.

To summarize: in an attempt to point out a few aspects that may explain some of the confusion surrounding the topic and – as a starting point for an analysis of regulative options – it can be argued that AI

- is a multifaceted *field of research and development (R&D),*
- generates and is dependent on specific *methods*,
- creates *applications* adapted for varying contexts,
- is what comes next.

2 What is Law?

Most people see laws as rules written on paper, defining matters, stipulating obligations and prohibitions, and being formally administrated by public authorities and courts. It is also known that laws should be issued in accordance with well-defined processes,[5] and, in order to uphold the rule of law,[6] be precise, clear and predictable. Altogether, the common understanding is that laws provide the scaffold for the legal sector of a society, which is divided into sub-sections such as criminal law, family law, IT law, administrative law, etc.

Such a description may give the impression that laws are disqualified as instruments to regulate a diversified, vague and dynamic

5 Usually according to procedures stipulated in constitutions, which are often laws of a special type, known to be of a general kind and protected by formalities, making them more resilient to hasty changes.

6 The rule of law here refers to the principle that the authority of law depends on the upholding of a number of (basically formal) principles. However, the concept has several interpretations in different jurisdictions.

phenomenon such as AI. In addition to the problems created by the fact that laws should be precise, clear and predictable, the inherent dynamics of AI applications makes it difficult to understand how the consequences of AI should be possible to foresee, let alone attribute to conceptions in existing laws. The multifaceted nature of AI, with the potential to affect all aspects of society, likewise seems to rule out the development of a specific legal sub-section. Nor is it readily conceivable that it is possible to develop one piece of legislation able to satisfy such broad demands. The question identified in the introduction thus reappears in a distinct form: if adequate laws are not available – must AI be regulated in new, alternative and/or complementary ways?

Before this question can be answered, it is necessary to elaborate the description of law and the legal sector. The understanding of the law as outlined above is not incorrect, but incomplete. Laws are not only static rules written on paper; as will be illustrated, they are also significantly transformative tools. Furthermore, it is important to recognize that laws and legal instruments have no inherent values of their own. Laws are correlated to and reflect the perceived needs of a society at a given point of time. From a functional perspective, laws are steering instruments designed and enacted in order to support political goals and manage problems. Laws exist because they have a purpose.

The law is not the same today as it was 10, 50, 100, 500 or 2,000 years ago. The changes have been fundamental, but one aspect is unquestionably stable: laws are flexible instruments. Throughout history, laws have been crucial for finding ways of managing problems and balancing conflicting interests of all types. The indisputable fact is that the history of law shows that almost everything has a legal side to it and can be legally regulated, and in this respect rapidly evolving technologies provide no exception.[7]

In order to explain this paradox, i.e., that precise, clear and predictable laws should be apt to regulate a diversified, vague and dynamic phenomena as AI, it is important to recognize that laws do not exist in a vacuum. Nor are laws uniform matters. In operative modes, laws must be studied in connection with other components,

7 Illustrations from recent centuries with huge impact on society include but are not limited to the development of legal regimes for railways, electricity, road traffic, air travel, nuclear power and telecommunications.

most importantly with so-called legal sources, of which legislative preparatory works, court decisions (case law) and scholarly commentaries are the most important.[8] Such complements create flexibility. Preparatory works offer information on underlying purposes and long-term goals, case law provides interpretational support via illustrations, and jurisprudential analysis adds complementary insights and recommendations concerning future developments.[9] In addition, a rich toolbox of legal methods provides a means to take contextual and dynamic aspects of regulated subject matters into consideration via interpretations, helping to preserve consistency of the legal system, in so far as possible.[10]

The ability to make regulations flexible is also supported by the fact that laws operate in parallel and at different levels, ideally creating a continuum of general and specific provisions, able to address issues from various perspectives. In legal analysis it is therefore necessary to make an inventory of relevant regulations and juxtapose legal components, indicating consequences relating to different aspects and prerequisites. Again, if the law is unclear on a specific matter, or regulative voids or clashes between poorly coordinated rules appear, interpretative methods have to be employed, which, if properly managed, are a guarantee for reaching decisions and a safeguard for upholding the rule of law.[11]

The fragmented nature of legal components is illustrated by the legal context of autonomous vehicles. Manufacturers of autonomous vehicles must *inter alia* respect national laws on traffic, vehicle classifications, security, patents, unfair competition, environment protection, insurance, liability, work force employment, taxes, etc. These laws operate horizontally in the sense that they all focus on

8 The hierarchical order between legal sources varies between jurisdictions and areas of law, and customs are sometimes included. In *common law* jurisdictions, case law is usually considered to have a higher status as compared to in *civil law* jurisdictions, in which legislation is predominant.

9 Sometimes referred to as arguments *de lege ferenda,* i.e., arguments concerning the law that is (or ought) to come into force.

10 *See*, for an overview with further references over methods for interpretation and legal reasoning, Wahlgren, P., *Legal Reasoning A Jurisprudential Model*, Scandinavian Studies in Law, Volume 40, Stockholm Institute for Scandinavian Law (2000).

11 A court cannot refrain from reaching a decision due to the fact that the law is silent on a certain matter, *jura novit curia*, "the court knows the law."

different aspects of the activity and must be observed in parallel. International projects accumulating vast collections of datasets are another illustration. Such projects – which are relevant for developers of autonomous vehicles in several ways – must in a similar way adhere to international laws and directives, for instance on data security, intellectual property rights, privacy, secrecy and freedom of information. Initiatives of this kind must however also take notice of hierarchical specifications stipulated in national laws, which in turn may be elaborated in public ordinances, court precedents and, in some cases, further detailed in conditional licenses issued by local agencies.[12] A manufacturer of autonomous vehicles must in this way observe both horizontally and hierarchically related provisions. For the legislator, on the other hand, regulative structures of this kind facilitate flexibility. When the need for changes occur, it is usually sufficient to revise one or a few components.[13]

Besides support from traditional legal sources, the dynamics of the legal system is enhanced by self- and co-regulating additions, collectively described as soft law. Soft law refers to recommendations, good practice standards, opinions, ethical principles, declarations, guidelines, board decisions, codes of conduct, negotiated agreements, private dispute resolution and a large number of additional normative mechanisms, united by the fact that they are results of non-state initiatives. Soft law is often initiated and promoted by

12 In Sweden, the *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC [GDPR]*, EUR-Lex – L:2016:119: is supplemented by the *Act containing supplementary provisions to the EU General Data Protection Regulation (SFS 2018:218)* and a large number of ordinances, with specific provisions for various activities (debt collection, patient data, police work, transports, etc.). In addition, in order to initiate certain research activities, it is necessary to apply for a permit, which may specify further conditions; *see* Etikprövningsmyndigheten (the Swedish Ethical Review Authority), https://etikprovningsmyndigheten.se/. *Cf.* also, The European Data Protection Board, *Guidelines 1/2019 on Codes of Conduct and Monitoring Bodies under Regulation 2016/679* Version 2.0, 4 June 2019, https://edpb.europa.eu/our-work-tools/our-documents/guidelines/guidelines-12019-codes-conduct-and-monitoring-bodies-0_en.

13 The vast majority of new laws are adjustments and changes of established laws. The Swedish Government's database with official and authentic versions of Acts and Ordinances lists 4,980 new entries for the period 1 April 2018–25 March 2021; of these, 4,821 (96.8%) have titles worded as "Act on change in act …" or "Act on change in ordinance …" https://svenskforfattningssamling.se/.

branch organizations, private enterprises, unions, etc. They can be intended for internal use, but sometimes have broader implications.[14] The concept of soft law can also be considered to include co-regulations in the form of traditional laws delegating regulative powers to private organizations, stipulating the setting up of domain-specific quality assurance systems open for inspection. The latter is an efficient way of creating flexibility and securing proficiency by means of involving stakeholders concerning domain-relevant subject matters.

Soft law norms are, and have more or less always been, continuously developed in areas where there is a need for regulation and the traditional law-making process is ineffective, time-consuming, difficult to manage or impractical. Frequently mentioned advantages are therefore that introduction of soft law solutions can be less time-consuming as compared to traditional legislation, that they require less bureaucracy and that they place less administrative burdens on the stakeholders. Illustrations can be found in various sectors of society, defining good practices in general or specifically, e.g., for medical treatments, or in the form of codes of conducts for media workers, lawyers and other professions. Soft law instruments often play an important role, especially in significantly technical domains where systems for component standardization, licensing, quality checks and accreditation of actors have been developed. It is also worth noting that the importance of soft law is in no way decreasing. International as well as national soft law are being recognized as important regulatory components, which indicates that the development of ethical codes and similar instruments for AI is a relevant path to investigate.

Soft law instruments are however seldom or never stand-alone features. They have to be tolerated and sometimes framed by the legislature. In some areas soft law solutions are impossible, e.g., because no suitable body with sufficient resources exists, when coordination requires joint efforts from many different stakeholders or if potential negative consequences may affect society as such.[15] Additional

14 Historically, the term soft law was initially used in order to denote international agreements, lacking operative mechanisms for ensuring compliance, but the concept is currently used in a broader sense. *See*, Wahlgren P. (ed.), *Soft Law*, Scandinavian Studies in Law, Volume 58, Stockholm Institute for Scandinavian Law (2013).

15 Security reasons largely exclude critical infrastructures and private initiatives potentially violating the law or threatening state interests.

limitations are that few or no effective sanctions exist in cases of breaches and that the lack of enforcement mechanisms may seriously disadvantage weaker parties. The absence of public channels for publication of internal codes of conduct as well as results of private conflict resolutions and arbitration can also create uncertainty about the content and significance of soft law jurisdictions.

3 Conclusion

As illustrated above, laws do not exist in isolation. Combined with contextually adjusted components, they form flexible regulative regimes of high complexity. In order to know to what extent ethical codes or other forms of regulations are required to manage AI, it is therefore necessary to develop an understanding of whether and to what extent the law is disqualified as a regulative mechanism for AI – developing new regulative systems without having access to a proper gap analysis is not a good idea.

Starting from the description of AI provided in section 1 above, this indicates that *if* AI is a field of R&D that generates new methods and dynamically operates applications able to adjust themselves to contextual requirements, *then* an analysis of regulative requirements related to AI must begin with an inventory of established legal doctrines and existing regulations concerning those matters (see Table).

Legal gap analysis of AI	
If AI	*Then* make an inventory of legal instruments able to regulate
is a multifaceted field of R&D	R&D
generates and is dependent on specific methods	methods
creates technical applications adapted for varying contexts	technical applications
is what comes next	dynamically

Table: Legal gap analysis of AI.

Contributors

Johan Axhamn

Assistant Professor in Business Law, Lund University

Johan Axhamn is an Assistant Professor in Business Law at Lund University. He obtained his LL.D. at the Faculty of Law at Stockholm University, in 2017, and also holds an LL.M. and a M.Sc. in Business and Economics from Lund University. Johan specializes in IT Law, Media Law, Market Law and Intellectual Property Law and related subjects. He has been a visiting scholar, inter alia, at Columbia University (New York), the University of Amsterdam, and the University of Oslo. Axhamn is Head of Unit at Lund University Centre for Business Law. He is also Editor-in-Chief of the Nordic Intellectual Property Law Review (NIR). Since 2012, Johan has served as a special adviser to the Swedish Government on intellectual property issues.

Liane Colonna

Assistant Professor in Law and IT, The Swedish Law and Informatics Research Institute, Department of Law, Stockholm University

Liane Colonna is currently employed as an Assistant Professor in law and information technology at the Department of Law, Stockholm University (SU) where she is performing research in the Wallenberg AI, Autonomous Systems and Software Program – Humanities and Society program. In particular, Liane is interested in ethical and legal challenges in relationship to AI-driven practices in higher education. She is also contributing to the development of privacy-aware video-based technologies and services for active and assisted living through her participation in several research projects on this topic. She has been a member of the New York Bar since 2008.

Joseph Early

PhD Student at the AIC Research Group, University of Southampton

Joseph Early is a doctoral student with the University of Southampton and the Alan Turing Institute. He completed an undergraduate degree in computer science at the University of Southampton, and then specialised with a master's degree focused on machine learning (also at the University of Southampton). Joseph pursed his interest in machine learning and is now in the third year of his PhD. His research focuses on interpretable machine learning – the process of unpacking the "black box" of artificial intelligence and understanding why models exhibit certain behaviours.

Martin Ebers

Associate Professor of IT Law at the University of Tartu, Permanent Research Fellow at the Humboldt University of Berlin

Martin Ebers is Associate Professor of IT Law at the University of Tartu, Estonia, and as "Privatdozent" permanent research fellow at the Humboldt University of Berlin, Germany. He is co-founder and president of the Robotics & AI Law Society (RAILS). In addition to research and teaching, he has been active in the field of legal consulting for many years. His main areas of expertise and research are IT law, liability and insurance law, and European and comparative law. In 2020, he published the books "Algorithms and Law" (Cambridge University Press), "Rechtshandbuch Künstliche Intelligenz und Robotik" (C.H. Beck) and "Algorithmic Governance and Governance of Algorithms" (Springer Nature).

Katarina Fast Lappalainen

Assistant Professor, The Swedish Law and Informatics Research Institute, Department of Law, Stockholm University

Katarina Fast Lappalainen is a Senior Lecturer (Assistant Professor) in Law and Information Technology at the Swedish Law and Informatics Research Institute, Law Department at Stockholm University. She has a J.D. in constitutional law and is a former tax lawyer. She has published articles and book chapters in different fields of law and has been the editor for several anthologies such as "AI and

Fundamental Rights". She has substantial experience teaching law in various fields at Stockholm University and is often engaged as a speaker at various seminars and conferences.

Sara Gandrén

Co-founder, Partner and Legal Advisor at inTechrity

Sara Gandrén is a privacy professional and entrepreneur. She holds a Swedish law degree and runs the compliance firm inTechrity, which supplies consultancy services mainly in the field of privacy and consumer protection. As such, she has been involved in a number of projects concerning artificial intelligence, consumer profiling and the Internet of Things. She enjoys doing research and writing about new technologies from a legal, philosophical and social perspective.

Stanley Greenstein

Assistant Professor in Law and IT, The Swedish Law and Informatics Research Institute, Department of Law, Stockholm University

Stanley Greenstein (LL.D.) is a Senior Lecturer (Assistant Professor) in Law and Information Technology at the Department of Law, Stockholm University. He is also a co-worker at the Swedish Law and Informatics Research Institute (IRI). Stanley's main area of interest is the interaction between digitalization and society. In this regard, his teaching, research and participation in project work has centered on the topic of artificial intelligence, machine learning and their ethical and societal implications. Stanley is also course director for the optional advanced course Cyber Law. A South African trained lawyer, Stanley has experience of working in a mixed legal jurisdiction made up of both the civil law and common law legal traditions.

Keri Grieman

Doctoral researcher and research associate, The Alan Turing Institute and Queen Mary University of London, University of Oxford

Keri Grieman is a research associate at the University of Oxford and a doctoral student at The Alan Turing Institute and Queen Mary University of London (QMUL). She holds a Master of Laws from QMUL, and a Juris Doctorate from the University of Calgary. Keri is a qualified lawyer in Ontario, Canada; was previously the Google

Policy Fellow at the Canadian Internet Policy and Public Interest Clinic; and was recently a consultant for The Ada Lovelace Institute on the proposed European Union Artificial Intelligence Regulation. Keri's research interests include responsible innovation and robotics, regulation of artificial intelligence, and practical applications of emerging technologies.

Håkan Hydén

Senior Professor in Sociology of Law, Lund University

Håkan Hydén is a Senior Professor in Sociology of Law at Lund University. Before that he was senior lecturer at the Department of Business Law and Docent in Private Law at the Law Faculty, Lund University. Håkan had the Chair in Sociology of Law between 1988 and 2012, and was appointed Samuel Pufendorf Professor 2008 until 2012. He is a fellow of the World Academy of Arts and Sciences since 2009, and fellow of Stellenbosch Institute for Advanced Studies since 2016. Håkan has been engaged in Sociology of Law on an international level by being for instance member of the governing board of the International Institute for Sociology of Law in Onati, Spain, and Vice President in the governing board of the Research Committee for Sociology of Law. He has served as a panel member in the European Research Council evaluating application for Consolidator Grants between 2016 and 2018. His main academic ambition is to consolidate the subject Sociology of Law as a Norm Science. His research interest is about how the technology affects society and law, how digital technology via AI in general and algorithms in particular form normativity in society.

Ubena John

Judge of the High Court of Tanzania

Ubena John is a Judge of the High Court of Tanzania, a Senior Lecturer and former dean of Faculty of Law Mzumbe, Tanzania. He has experience in legal practice, legal training, conducting research and consultancy and supervising research works for undergraduate and postgraduate law students. He has taught law courses for about sixteen years. Ubena has been teaching ICT Law since 2008. Ubena is a member of the African Union Cybersecurity Expert Group (AUCSEG). He holds LL.B. from Mzumbe University, LL.M (Law

and Information Technology) and LL.D. in Information Technology Law from Stockholm University. He has conducted research projects focusing primarily on ICT Law. Moreover, he has undertaken consultancy assignments for both private and public institutions, as well as international corporations. He has participated in among other things drafting of Tanzania Judicature and Application of Laws (Electronic Filing) Rules, 2018. Furthermore, Ubena has authored several scholarly works on ICT Law, cybercrimes legislation, privacy, legislative techniques and regulatory law.

Nicklas Berild Lundblad

Ph.D. in Applied Information Technology, Head of Global Tech Policy at Stripe

Nicklas Berild Lundblad is an interdisciplinary scholar who merges tech, society and future studies. He has held different positions with Google since 2007, working for the company in a number of roles for 13 years, leaving in 2020. He now works at Stripe, leading their EMEA and APAC policy as well as their tech policy work. He holds a PhD in informatics, a Swedish law degree, and a BA in philosophy. He is a frequent contributor to magazines and newspapers, writing about cognitive science, law, future studies, artificial intelligence and philosophy. He recently published a book on the philosophy of questions.

Cecilia Magnusson Sjöberg

Professor of Law and IT, The Swedish Law and Informatics Research Institute, Department of Law, Stockholm University

Professor Cecilia Magnusson Sjöberg is Subject Director of Law and Information technology at Stockholm University. She was awarded a LL.D. degree in 1992, with a doctoral thesis addressing legal automation, especially about the computerisation in public administration. Legal implications of e-government remains one of her major fields of work. In addition to substantive components of IT Law, e.g. privacy protection, she has had many years of experience of legal system design and management, giving rise to information security issues and the need for electronic signatures etc. In addition to a wide variety of national and international research projects addressing the interplay between law and modern information communication

technologies she is engaged by the Swedish government in public inquires about e.g. personal data protection for research purposes and how to legally facilitate the digitalisation of the public sector.

Tobias Mahler

Professor, Norwegian Research Center for Computers and Law (NRCCL), University of Oslo

Tobias Mahler teaches law at the faculty of law at the University of Oslo. He is a professor at the Norwegian Research Center for Computers and Law (NRCCL), specializing in information and communications technology law. His research interests cover a broad range of legal issues arising in the context of (i) robots, particularly with artificial intelligence capabilities, (ii) Internet governance (especially the domain name system), as well as (iii) cybersecurity and privacy. This focus on legal issues is complemented with research interests in legal informatics more closely related to computer science. The latter line of research has focused on software applications for legal practice, such as, legal risk management and visual representations of legal reasoning. He holds a PhD from the University of Oslo, an LLM degree in legal informatics from the University of Hannover, and a German law degree (first state exam). He has practised law in Norway as corporate lawyer in the automotive industry, primarily working with international commercial contracts. He teaches primarily robot regulation, cybersecurity regulation, legal tech and artificial intelligence. He is the deputy director of the NRCCL and the director of the centre's LLM programme.

Chris Reed

Professor of Electronic Commerce Law at the Centre for Commercial Law Studies, Queen Mary University of London

Chris Reed is Professor of Electronic Commerce Law at the Centre for Commercial Law Studies, Queen Mary University of London, where he was formerly Director of the Centre and subsequently Academic Dean of the Faculty of Law & Social Science. He consults to companies and has worked exclusively in the computing and technology law field since 1987. He also teaches University of London LL.M. students from all over the world. Chris has published widely on many aspects of computer law, and research with which he was

involved led to the EU directives on electronic signatures and on electronic commerce. The Leverhulme Foundation awarded him a Major Research Fellowship for 2009–2011. From 1997 to 2000 Chris was Joint Chairman of the Society for Computers and Law, of which he is an inaugural Honorary Fellow, and in 1997–1998 he acted as Specialist Adviser to the House of Lords Select Committee on Science and Technology. Chris has acted as an Expert for the European Commission, represented the UK Government at the Hague Conference on Private International Law and has been an invited speaker at OECD and G8 international conferences.

Richard Sannerholm

Senior Lecturer, Institution for Social Sciences, Södertörn University

Richard Sannerholm is a Doctor of Law and a Senior Lecturer at Södertörn University. He is an editor with the Hague Journal on the Rule of Law and a member of the Swedish National Commission for UNESCO. Richard writes about rule of law and public administration and has worked extensively on rule of law reform in conflict and crisis settings. His latest book was on the rule of law in Sweden – Rättsstaten Sverige. Skandaler, kriser, politik (2020).

Caroline Sundberg

LL.M and Member of the Swedish Bar Association, Attorney at Hannes Snellman Attorneys Ltd

Caroline Sundberg is an attorney (LL.M and Member of the Swedish Bar Association) at Hannes Snellman, where she is responsible for the Data and Cyber Security team at the Stockholm office. In her practice she is specialised in the law related to the IT and technology, with a particular focus on commercial agreements, cyber security and data privacy (GDPR). She regularly advises clients, from both the public as well as the private sector in this field. Caroline has during the last decade obtained considerable experience of matters related to evolving technologies and has during recent years been involved in several AI-related projects.

Jessica Tressfeldt

LL.M and Associate at Hannes Snellman Attorneys Ltd

Jessica Tressfeldt is part of Hannes Snellman's IP & Tech team. She works mainly within the field of IT, technology and data privacy (GDPR).

Katja de Vries

Assistant Professor in Public Law, Uppsala University

Katja de Vries is an Assistant Professor in public law at Uppsala University funded by the Ragnar Söderberg Foundation. She is also affiliated to the Swedish Law and Informatics Research Institute (Stockholm) and the Center for Law, Science, Technology and Society (Brussels). Her current research focuses on the challenges that AI-generated content ('deepfakes' or 'synthetic data') poses to data protection, intellectual property and other fields of law.

Peter Wahlgren

Professor of Law and IT, The Swedish Law and Informatics Research Institute, Department of Law, Stockholm University

Peter Wahlgren is Professor of Law and Information Technology. He is currently Torsten and Ragnar Söderberg Professor of Legal Science and is chairman and Director of IRI. He was awarded the degree LL.D. in 1992, Automation of Legal Reasoning: A Study on Artificial Intelligence and Law (Kluwer). Docent in Jurisprudence (Allmän rättslära), Docent in Law and IT (Rättsinformatik) and appointed professor in Law and IT in 2001. His research interests cover automated legal methods, IT/AI and law, proactive law, legal risk analysis and legislative techniques.

Nordic Yearbook of Law and Informatics 2020–2021

Law in the Era of Artificial Intelligence

Law in the Era of Artificial Intelligence includes a collection of papers presented at the 35th Nordic Conference on Law and Information Technology. The Conference was held in Stockholm in November 2020 and had the title 'Law in the Era of Artificial Intelligence'. It examined the manner in which legal concepts, developed for a non-digital context, can continue to be applied in the era of artificial intelligence. The Nordic Yearbook of Law and Informatics 2020–2021 explores the above theme and is divided into four parts: data protection, transparency, liability and regulation.

The Nordic Yearbook of Law and Informatics (Nordisk årsbok i rättsinformatik) was initiated by the Scandinavian institutes for Law and Informatics. It was first published in 1984 and during the initial period a yearbook was produced on an annual basis. From its inception, hundreds of articles have been published, reflecting developments in the Scandinavian context and internationally.

The Swedish Law and Informatics Research Institute (IRI) is a research unit at the Faculty of Law, Stockholm University, and was founded in 1968. Having been instrumental in their establishment, IRI has had a long working relationship with The Swedish Society for IT and Law (Svenska föreningen för IT och juridik, SIJU) and The Foundation for Legal Information (Stiftelsen för rättsinformation), operating since 1988.

Distributor: eddy.se ab
www.bokorder.se
order@bokorder.se

Produced by Stiftelsen Juridisk Fakultetslitteratur (SJF)
and The Swedish Law and Informatics Research Institute (IRI)

ISBN: 978-91-8892-964-8